ASME PTB-12-2017

Guidelines for Addressing Data Gaps and Recordkeeping for ASME B31.4, B31.8 and B31.8S for Pipeline Integrity Management

ASME
SETTING THE STANDARD

PTB-12-2017

GUIDELINES FOR ADDRESSING DATA GAPS AND RECORDKEEPING FOR ASME B31.4, B31.8 AND B31.8S FOR PIPELINE INTEGRITY MANAGEMENT

Prepared by:

Andrew R. Lutz, P.E., DNV-GL
Elizabeth King, DNV-GL
Dr. Thomas Bubenik, DNV-GL
Dr. William Harper, DNV-GL

ASME
SETTING THE STANDARD

Date of Issuance: June 28, 2017

This guideline was prepared by ASME Standards Technology, LLC (ASME ST-LLC) and sponsored by The American Society of Mechanical Engineers (ASME) Pressure Technology Codes & Standards.

Neither ASME, the author, nor others involved in the preparation or review of this document, nor any of their respective employees, members or persons acting on their behalf, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe upon privately owned rights.

Reference herein to any specific commercial product, process or service by trade name, trademark, manufacturer or otherwise does not necessarily constitute or imply its endorsement, recommendation or favoring by ASME or others involved in the preparation or review of this document, or any agency thereof. The views and opinions of the authors, contributors and reviewers of the document expressed herein do not necessarily reflect those of ASME or others involved in the preparation or review of this document, or any agency thereof.

ASME does not "approve," "rate," or "endorse" any item, construction, proprietary device, or activity.

ASME does not take any position with respect to the validity of any patent rights asserted in connection with any items mentioned in this document, and does not undertake to insure anyone utilizing a standard against liability for infringement of any applicable letters patent, nor assume any such liability. Users of a code or standard are expressly advised that determination of the validity of any such patent rights, and the risk of infringement of such rights, is entirely their own responsibility.

Participation by federal agency representative(s) or person(s) affiliated with industry is not to be interpreted as government or industry endorsement of this code or standard.

ASME is the registered trademark of The American Society of Mechanical Engineers.

The American Society of Mechanical Engineers
Two Park Avenue, New York, NY 10016-5990
ISBN No. 978-0-7918-7171-3

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

FOREWORD

The San Bruno event on September 9, 2010 was paradigm shifting for the pipeline industry. This guideline represents a continuing effort to learn from our mistakes, to learn from outside sources, to share knowledge and to improve the pipeline industry in the interest of public safety.

The following individuals are acknowledged for their technical and editorial peer review of this guideline: David Anderson, Michael Rosenfeld, Marvin Hovis, Joel Brandt, Keith Leewis, David Johnson, Michael Zerella, and Rick Kivela. In addition, the efforts of Richard Lucas of ASME and Carlton Ramcharran of ASME ST-LLC are acknowledged for their management of the peer review group, review of the manuscript prior to publication, editing and document preparation resulting in the publication of this document. Finally, a special thanks is offered to the pipeline operators who invested in developing processes and technologies over the past five years that supported the development of this document.

Established in 1880, ASME is a professional not-for-profit organization with more than 135,000 members and volunteers promoting the art, science and practice of mechanical and multidisciplinary engineering and allied sciences. ASME develops codes and standards that enhance public safety, and provides lifelong learning and technical exchange opportunities benefiting the engineering and technology community. Visit (https://www.asme.org/) for more information.

ASME ST-LLC is a not-for-profit Limited Liability Company, with ASME as the sole member, formed in 2004 to carry out work related to new and developing technology. The ASME ST-LLC mission includes meeting the needs of industry and government by providing new standards-related products and services, which advance the application of emerging and newly commercialized science and technology, and providing the research and technology development needed to establish and maintain the technical relevance of codes and standards. Visit (http://asmestllc.org/) for more information.

ABSTRACT

This guideline provides a recommendation and potential guidance to address observed gaps in data and recordkeeping practices that are currently prescribed in the ASME pipeline standards B31.8, Gas Transmission and Distribution Piping Systems, B31.8S, Managing System Integrity of Gas Pipelines, and B31.4, Pipeline Transportation Systems for Liquid Hydrocarbons and Other Liquids. The authors reviewed pipeline industry standards and standards from other industries to identify potential best practices and lessons learned.

In practice, once operators have established upgraded data and recordkeeping systems, the systems will need to be managed and maintained, or else there would be risk of repeating old mistakes, and the need to re-create the system from the ground up. A records maintenance team and controlled processes are required to manage and maintain the systems effectively. This guideline endeavors to summarize guidance that a pipeline operator could use to enhance a data and recordkeeping structure in accordance with modern standards.

It is recommended that each operator maintain a chief source of pipeline system information, so he can make auditable, repeatable, and trustworthy decisions such as those needed for fitness-for-service calculations and risk assessments. The authors of this guideline recommend a practice of maintaining digital source records directly linked to specific pipeline components within an operator's geographic information system (GIS) (i.e., a pipeline information database). This recommendation is based on a philosophy of providing the operator's decision makers easier access to the source records.

Guidance is provided on bounding likely values to address data gaps through research. The purpose of the research is to gather information that allows the team to assign conservative, realistic ranges of values for missing parameters.

Guidance is provided to develop a quality and reliability process for pipeline system data. If the existing quality and reliability determination process is found unacceptable or it is found that one does not exist, then a new process must be created. A team should be assembled to research information to accurately develop a quality and reliability determination process. If the existing quality and reliability determination process is found to be acceptable, then it should be communicated to relevant personnel. This process is to be adopted and will be used to verify the pipeline information database.

Data collection, as it pertains to information about a pipeline system, has always occurred continuously in the pipeline industry. Operators and their contractors are constantly recording observations and documenting measurements as part of field, survey, and testing activities. These new observations and measurements in effect create new source records with respect to the pipeline information database. An ongoing data collection process is described to maximize the utility of this data towards filling gaps, increasing confidence, and maintaining the database into the future.

ABBREVIATIONS AND ACRONYMS

Acronym / Abbreviation	Meaning
ADB	Advisory Bulletin
AGA	American Gas Association
ANSI	American National Standards Institute
API	American Petroleum Institute
ASME	American Society of Mechanical Engineers
APDM	ArcGIS Pipeline Data Model
CFR	U.S. Code of Federal Regulations
DOT	U. S. Department of Transportation
DPI	Dots per inch
FAQ	Frequently Asked Questions
FDA	U.S. Food and Drug Administration
GIS	Geographic Information System
GPS	Global Positioning System
HCA	High Consequence Area
IMP	Integrity Management Program
ISO	International Organization for Standardization
MAOP	Maximum Allowable Operating Pressure
MOC	Management of Change
MOP	Maximum Operating Pressure
NACE	NACE International (formerly) National Association of Corrosion Engineers
NPMS	National Pipelines Mapping System
NTSB	U.S. National Transportation Safety Board
O&M	Operations and Maintenance
PDF	Portable Document Format
PHMSA	U.S. Pipeline and Hazardous Materials Safety Administration
PODS	Pipeline Open Data Standard
PPI	Pixels per Inch
PRCI	Pipeline Research Council International
QA/QC	Quality Assurance/ Quality Control
SCADA	Supervisory Control and Data Acquisition
SME	Subject Matter Expert
SMYS	Specified Minimum Yield Strength
UPDM	Utility and Pipeline Data Model

DEFINITIONS

1 Access: right, opportunity, means of finding, using, or retrieving information
2 Bayesian Network: graphic statistical model that represents probabilistic relationships between variables. This type of analytic process can model the probabilities of cause and effect relationships in order to make predictions based on known distributions and values (see Appendix A).
3 Company segment: A natural division within a company that may drive a separate document structure, such as a legacy acquisition, or geographic region (e.g., "southwest district") where a unique document organization may exist.
4 Complete record: a record that is "finalized by a signature, data, or other appropriate marking."
5 Conversion: process of changing records from one format to another.
6 Data Gap: missing information.
7 Data mining: the practice of searching through large amounts of information, to locate specific data or identify useful trends.
8 Database: a large collection of data organized for quick access and retrievability.
9 Database Owner: a person responsible for managing and maintaining the data within the Pipeline Database for the relevant section of pipe.
10 Decision Maker: The staff and stakeholders whose decisions require the information from the Pipeline Database, which may include critical operations decisions (e.g., pressure reductions and fitness-for-service analysis). Examples of Decision Makers include company management, integrity engineers, and risk engineers.
11 Destruction: process of eliminating or deleting a record, beyond any possible reconstruction.
12 Disposition: range of processes associated with implementing records retention, destruction, or transfer decisions which are documented in disposition authorities or other instruments.
13 Flowchart: a graphical representation of a process.
14 Media: the physical form of the documentation such as paper, or electronic.
15 Metadata for records: structured information which enables the indexing, sorting, retrieval, and use of records.
16 Metric: a standard of measurement by which quality, performance, or progress may be measured.
17 Microfiche: a flat piece of film containing microphotographs of the pages of a newspaper, catalog, or other document. [1]
18 Microfilm: a length of film containing microphotographs of a newspaper, catalog, or other document. [1]
19 Other Communications (pertaining to regulatory requirements): U.S. Pipeline and Hazardous Materials Safety Administration (PHMSA) Emails, Letters, Advisory Bulletins, and FAQs.
20 Pipeline Information Database: an organized collection of data specific to the needs and decisions of the pipeline company.
21 Record(s): Recorded information or data on a particular subject, collected and preserved to demonstrate compliance with a rule or process requirement. [2]
22 Records Librarian: a person who is familiar with the hard copy records pertaining to the data that are or will be used to populate the Pipeline Database.
23 Records management: field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use, and disposition of records, including processes for capturing and maintaining evidence of, and information about, activities in the form of records.

24 Records system: information system which captures, manages, and provides access to records over time.
25 Researcher: a person who is available to conduct the necessary research into the background information related to the desired information.
26 Schema: logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax, and the optionality (obligation level) of values. [3]
27 SCADA: Supervisory control and data acquisition system, used for monitoring and control.
28 Source Records: a record that is original for the data included within.
29 Subject Matter Expert (SME): An individual recognized as having a special skill or specialized knowledge of a process in a particular field, or of a piece of equipment. [4] Types of SMEs include welding and materials experts, pipeline construction project experts, legacy acquired company experts, etc.
30 Traceable record: a record that can clearly be linked to original information about a pipeline segment or facility.
31 Useable: (pertaining to records) a record/data that is accessible to the Decision Maker within a reasonable time period. [5]
32 Verifiable record: a record that contains information that can be "confirmed by other complementary, but separate, documentation."

1 PURPOSE AND USE

This guideline provides a recommendation and potential guidance to address observed gaps in data and recordkeeping practices that are currently prescribed in the ASME pipeline standards B31.8, B31.8S, and B31.4. The authors reviewed pipeline industry standards and standards from other industries to identify potential best practices and lessons learned. The industry needs a go-forward approach that will manage new records and data in a way that will maximize the utility of available records wherever practical, and improve upon record keeping practices for the future to take advantage of modern technology and lessons learned.

In practice, once operators have established upgraded data and recordkeeping systems, the systems will need to be managed and maintained, or else there would be risk of repeating old mistakes and the need to re-create the system from the ground up. A records maintenance team and controlled processes are required to manage and maintain the systems effectively. This guideline summarizes guidance that a pipeline operator could use to enhance a data and recordkeeping structure in accordance with modern standards.

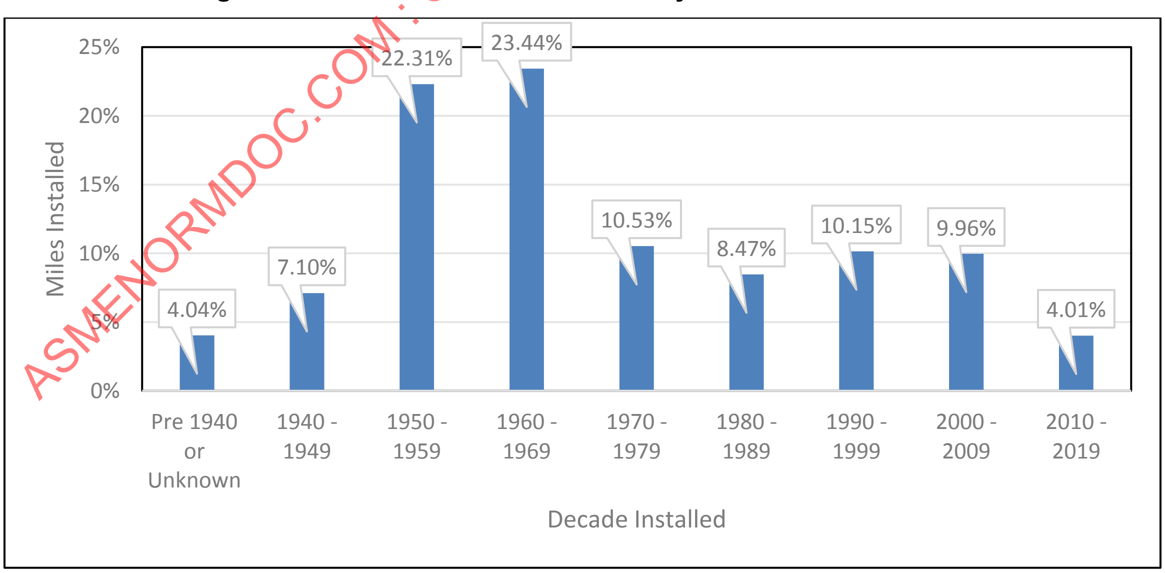
1.1 Background of Data and Recordkeeping in the Pipeline Industry

The pipeline industry's data and recordkeeping systems are a product of the industry's historical practices. In order to identify ways in which ASME codes and standards can provide guidance to industry, it is important to first understand historic pipeline record keeping practices and to understand how the industry's infrastructure and regulatory history steered the records to their current state.

1.1.1 United States

In the United States (U.S.), over one half of the transmission pipeline network in the country was installed prior to 1970 (and prior to any federal pipeline safety regulations with records requirements). Figure 1-1-1 shows gas transmission pipeline installation activity by decade, which peaked in the 1950s (22.31%) and the 1960s (23.44%). [6]

Figure 1-1: Gas Transmission Lines by Decade of Installment



Prior to the first U.S. pipeline regulations in late 1960s, the industry's only record keeping requirements were provided by industry standards. Rosenfeld and Gailing summarized the situation well, "It would be reasonable to expect that a variety of documents related to the design and construction of a pipeline facility be retained long-term. However, retention of technical documents was not addressed by the engineering standards of the day. It was generally thought that a copy of the specifications under which the pipeline was built (and supplemented by commercial documents, e.g. contracts and purchase orders) would generally be adequate to provide evidence of the work that was done." [7]

The authors have made similar observations during their professional experiences with records review and management. There are occasions where operators have stored what are now recognized to be highly valuable records, but it is highly unusual that they stored them to meet any specific regulatory requirement. When highly valuable documents are discovered (particularly for pipe that was constructed prior to regulation and/or pipe that was acquired), they are frequently met with pleasant surprise.

In 1938, American Standards Association ASA B31.1 first required that records be kept on welder qualifications and their identifying marks. Subsequent revisions expanded welder related record keeping. In 1955, B31.1.8 (which eventually became B31.8) first recommended basic risk based design concepts with 4 location class factors. It required the pipeline operator, or contractor, to maintain records related to welders and pressure testing. The standard was the first to recommend operations and maintenance records mentioning external and internal corrosion related to leaks and repairs, and inspection reports. A later revision in 1968 required recordkeeping related to corrosion inspection and leak investigation.

The federal pipeline regulations were passed in the 1960s with the first federal laws effecting liquid pipelines and the Natural Gas Pipeline Safety Act of 1968. Concurrent to these shifts in pipeline records requirements and expansion of the U.S. pipeline infrastructure indicated above was large amounts of population growth over the past 50-60 years. Pipelines had to be re-routed to accommodate the additional infrastructure (e.g., highways, waterlines) in congested areas, which created more records and/or additional pipeline system materials to track.

The Natural Gas Integrity Management Rule 49 CFR Part 193 Subpart O was introduced in 2003, three years after a similar rule for liquid pipelines (49 CFR Part 195). Following the integrity management rules, operators were more frequently audited, which required them to focus more on material properties as part of risk modeling and fitness-for-service analysis. The gas (and, similarly, the liquid) integrity management (IM) rule specified how pipeline operators must identify, prioritize, assess, evaluate, repair, and validate the integrity of gas (or liquid) transmission pipelines that could, in the event of a leak or failure, affect High Consequence Areas (HCAs) within the U.S. The IM rules required large improvement in HCA-related record keeping for most operators in the U.S. [8]

The industry heightened its focus on data and recordkeeping following the San Bruno failure in 2010, when the U.S. National Transportation Safety Board (NTSB) included in its findings that the pipe had been incorrectly listed as "seamless." The first records quality criteria in the U.S. were provided in 2011, with PHMSA's issuance of Advisory Bulletin (ADB) ADB-11-01 [15], which made operators aware that operational decisions should be based on documents that are "traceable, verifiable, and complete" (PHMSA provided definitions for the terms approximately 16 months later in ADB-12-06) [16]. Many operators had to satisfy this requirement[1] by locating, sorting, deciphering, and prioritizing decades of legacy and historic records. It was found that many records, particularly pre-regulation, were not retained as there were no requirements to keep them on file. Economic shifts have caused mergers and acquisitions (creating situations where the records may not have been transferred from one operator to the next). There are many other reasons for loss of records, including lack of retention of hardcopy after poor quality archiving to

[1] The requirement to review system records and verify MAOP was made into law by the U.S. Congress when the Pipeline Safety, Regulatory Certainty, and Job Creation Act of 2011 was passed on January 3, 2012. [17]

microfilm or digitization. Disasters such as a fire or flood, or clerical mishandling[2] have also destroyed invaluable records that are costly to reproduce. Operators of these pipelines have been placed in a difficult situation.

1.1.2 CSA Z662

Canadian Standard Association (CSA) standard Z662-15 provides a more flexible view of historical records with its requirements: "The format and level of detail associated with the material records specified in Clause 5.7 is not defined and is therefore at the discretion of the company. Although detailed documentation, such as mill test reports, often provide useful historical data for future reference (e.g., for engineering assessments pertaining to failure investigations or development of in-service welding procedures), it is not mandatory that such detailed documentation be retained as part of the permanent record. It is the intent that basic material data, such as material standards, specifications, grades, and dimensions, at a minimum, be included in the permanent records."

While the requirements of CSA Z662-15 may be a bit more flexible with respect to management of the pre-regulation era records, it allows operators to discard source records and doesn't require operators to manage their data with any type of reliability or quality criteria.

1.1.3 AS 2885

The Australian Standard 2885 series of standards (AS 2885.0 through AS 2885.5) [9][10][11][12][13][14] include some extensive lists of specific records requirements. For example, AS 2885.5 – 2012 (Field Pressure Testing) [14] contains approximately fifty specific and itemized reporting requirements for each test. It also provides an example of an acceptable format for a pressure test "certificate" that serves to record the required information.

The AS 2885 standards also require a Records Management Plan, and AS 2885.1 – 2012 (Design and Construction) [10] states "all pressure-containing materials installed on a pipeline system shall be traceable to the purchase documentation, the manufacturing Standard, the testing standard, and to inspection and acceptance documents. The pipeline Licensee shall maintain the records until the pipeline is abandoned or removed."

The Australian Design and Construction standard (AS 2885.1 – 2012) makes recommendations about digitization of records, as follows: "Electronic records that can be accessed by common text, database or spreadsheet programs are preferred. Where documents are only available on paper, they should be scanned into an appropriate format." And the standard requires the following link between materials and source records: "The identity of all materials shall be recorded and this identity shall include reference to the test certificates and/or inspection reports."

1.1.4 International

The International Organization for Standardization (ISO) standard 13623 [18] prescribes generally that "records shall be kept and maintained throughout (the pipeline's) lifetime to demonstrate compliance with the requirements of this international standard" and includes brief statements that require:

- Leak detection surveys.
- Records to demonstrate the system is "operated and maintained in accordance" to operations and maintenance and integrity plans and they are "effective."
- Pre-commissioning and commissioning records.

[2] "An anecdote reported to [Rosenfeld and Gailing] was an occasion where a clerical worker, instructed to photocopy hydrostatic test records, first separated the pressure charts from the test report forms which had been stapled together into separate piles, irreversibly breaking the link between pressure records and test segments." [7]

- Pressure test records, with specific itemized requirements including testing procedure, instrument calibrations, test charts, explanation and disposition of pressure discontinuities, etc.
- For construction records, it specifies some general document types. It prescribes a specific requirement that these documents should be made "permanent in reproducible and retrievable form," which seems to imply they should be available upon demand if someone needs them for an audit or analysis.

While the ISO standards are a bit more specific in some areas (e.g., pre-commissioning/commissioning), and less specific in others (e.g., operations and maintenance), the standard doesn't provide much records management guidance except for the last bullet related to construction records. The ISO standard recognized a specific need to continuously have access to the construction records. This improvement in practice of providing more access to records is also recognized by the nuclear industry in the next section.

NACE International (NACE) standard SP0113-2013 "Pipeline Integrity Method Selection" [19] provides an overview of available assessment methods and selection guidance to choose an appropriate method to assess the integrity of liquid or gas pipelines for external corrosion, internal corrosion, and SCC threats. This standard is primarily focused on records collection and prescribes very little (if any) records management. It does not prescribe or require any records from the selection process to be stored in any specific manner or for any specific amount of time. The NACE standard describes selecting a second assessment based on information from the first but doesn't describe how to store the information in between.

The reason the NACE document doesn't prescribe record keeping practices is because it is intended to complement a robust integrity management and operations and maintenance program, which should contain recordkeeping requirements. The NACE document provides technical guidance, but it may not be able to serve its purpose without adequate recordkeeping guidance provided by another standard.

1.1.5 Recent Improvements

The industry will need a go-forward approach that will manage new records and data in a way that will maximize the utility of available records as much as practical, and improve upon record keeping practices for the future, to take advantage of modern technology and lessons learned.

Modern records management programs have been presented at various U.S. pipeline operator conferences and forums frequently since around 2012. Operators in the U.S. have taken action in response to the NTSB's, PHMSA's, and U.S. Congressional recommendations and rules. Presenters across the industry have valued peer review of their data and records management review programs. Common themes of the review programs include organizing and searching source records, creating a document ranking system, and creating a cross-reference between the source records and the verified pipeline system materials data.

Some operators have found that digitization has allowed them to create a system that is more sustainable and allows more of the operators' employees to access data and recordkeeping system, which is perceived as an increase in value on the investment. Specific examples include operators that have digitally captured or transferred source records into a records management system, which allows the operator to initially conduct research efficiently, but ultimately links source records to GIS. The GIS system provides an interface with the pipeline records system that allows access for a much wider set of employees. These GIS-based systems can provide more accessibility than most traditional analogue systems (e.g.., like maps, or alignment sheets) because they can often be navigated more quickly and intuitively and on a system-wide scale. They can also provide more accessibility than a spreadsheet based cross-reference between source records and verified data because of the barriers to understanding the spreadsheet cross-reference.

Once established, operators have understood these upgraded data and recordkeeping systems will need to be managed and maintained, or else there would be need to re-create the system from the ground up. A records maintenance team and controlled processes are required to manage and maintain the systems effectively.

This guideline endeavors to summarize guidance that a pipeline operator could use to enhance a data and recordkeeping structure in accordance with modern standards.

1.2 Review of Data Management Challenges and Solutions in Other Industries

Comparisons of the general data and recordkeeping practices of the pipeline industry to other industries have highlighted some key observations. The pipeline industry has the following unique combination of characteristics that create challenges for data and recordkeeping. In the U.S., the pipeline infrastructure is:

- By majority more than 50 years old.
- Materials dependent (i.e. subject to degradation mechanisms like corrosion and cracking).
- Typically buried underground (and can't be readily seen or inspected).
- Vastly spread out geographically, and
- Highly critical to society's safety, environmental, and infrastructure (e.g., energy, economy) requirements.

The other industries reviewed for this guideline exhibit some of the same characteristics, but the authors couldn't identify an industry which exhibited all of these same characteristics. The authors reviewed practices within the following industries to evaluate how the recordkeeping practices might be transferred to the pipeline industry.

1.2.1 Medical Industry

Recordkeeping practices were reviewed for the implanted medical device sector of the medical industry. This part of the medical industry shares the following characteristics with the pipeline industry:

- Critical to safety;
- Can't be readily seen or inspected (once installed); and are
- Materials dependent.

Tracking requirements and practices are quite rigid within the U.S. Code of Federal Regulations (CFR) for implanted medical devices. Specifically, the medical industry requires a unique device number assigned to all medical devices implanted within patients, so the manufacturer can retrieve manuals, install dates, etc. for any of its devices for the U.S. Food and Drug Administration (FDA) within ten days of any request. [20]

This level of tracking is beyond what is required for most of the pipeline industry. For example, the marking and naming requirement is similar to what is required for valves [21] (where a unique id, nameplate, and body markings are required), but the requirements for how the information will be tracked after the valve is installed is currently very general.

The pipeline industry could benefit from a practice of uniquely tracking each component that is installed (although it would be tedious) and storing the information in a retrievable manner. This practice is more realistic as a future goal, since it may not be practical to retroactively obtain this level of detail from project records (i.e., from past or current projects).

1.2.2 Nuclear Industry

Recordkeeping practices were also reviewed for the nuclear industry. The nuclear industry shares the following characteristics with the pipeline industry:

- Can't be readily seen or inspected (once installed);
- Materials dependent; and
- Highly critical to society's safety, environmental, and infrastructure (e.g., energy, economy) requirements.

The nuclear industry in general was reviewed with respect to the recordkeeping requirements for comparison with the pipeline industry. There were a number of key findings during this review. The nuclear industry has a substantial number of robust practices related to data and recordkeeping requirements. Quality Assurance Requirements for Nuclear Facility Applications [22] prescribes requirements for:

- Identification of traceability of items;
- Authentication of records;
- Receipt control of records; and
- Maintenance of records.

Some of the requirements, like "Identification of traceability of items," specify that a heat number may be used to trace a material grade, but they don't elaborate with specific guidance on how this should be performed (e.g., within documentation).

"Authentication of records" serves to set the expectations of a source record (i.e., an "authentic data source"). For example, "Documents shall be considered valid records only if stamped, initialed, or signed and dated by authorized personnel or otherwise authenticated. Corrections to documents shall be reviewed and approved by the responsible individual from the originating or authorized organization."

These criteria are similar to the types of markings that often appear in pipeline industry construction documents. The authors of this guideline have reviewed construction and maintenance documentation from the pipeline industry and subjected it to similar criteria on a project-to-project basis. Although many operators have procedures, and forms that require signatures, dates, etc., the first requirement for all records to have such markings was PHMSA's advisory bulletin ADB-11-01 [15], which required pipeline operators to have "complete" records to verify their operating pressures.

The Nuclear Standard section on "receipt control of records" designates a role for receiving records, with the responsibility and organizing them in both temporary and permanent storage. A requirement for dedicated records personnel is one that many pipeline operators have begun to realize is needed. Operators have developed teams to "review" records initially, and they have maintained the teams as permanent records management fixtures once the initial review is complete.

The "maintenance of records" has requirements about accessibility, management of change, and version control of records, but it doesn't go into detail about how this will be completed.

In the ASME Boiler and Pressure Vessel Code Section III Division 3 (BPVC-III-D3) "Containments for Transportation & Storage of Spent Nuclear Fuel & High Level Radioactive Material & Waste," [23] there are also requirements that the pipeline industry could potentially learn from.

This standard also provides itemized tables of lifetime quality assurance records and nonpermanent quality assurance records, which are as follows in Figure 1-2.

Figure 1-2: Prescriptive Quality Assurance Records Requirements from BPVC-III-D3

Table WA-4134.17-1
Lifetime Quality Assurance Records

Record	Record
1. Index to lifetime records (NCA-4134.17)	9. Final nondestructive examination reports
2. Code Data Reports (WA-8400)	10. Repair records when required by Code (Article WB-4000, Article WC-4000)
3. Design Specification (WA-3300)	11. Weld procedures
4. Design Documents (WA-3300)	12. Audit and survey reports (NCA-4134.18)
5. As-built drawings (WA-3300)	13. Process sheets, travelers, or checklists
6. Certified Material Test Reports (CMTR) and documentation providing traceability to location used, if required (WB-4100, WC-4100)	14. Joint-welder identification records when such records are used in lieu of physical marking of welds (WB-4300, WC-4300)
7. Heat treatment records [Note (1)]	15. Fabrication Specification (WA-3300)
8. Final hydrostatic and pneumatic test results (Article WB-6000, Article WC-6000)	16. Casting Plan (WB-2126, WC-2126)

GENERAL NOTE: Nonconformance reports that affect those records listed shall be incorporated into the record or be retained with the records.

NOTE:
(1) Either heat treatment charts or certified summaries of time and temperature data may be provided. These data may be included as part of the CMTR.

Table WA-4134.17-2
Nonpermanent Quality Assurance Records

Record	Retention Period
1. QA Program Manual	3 yr after superseded or invalidated
2. Design procurement and QA procedures (NCA-4134.5)	3 yr after superseded or invalidated
3. NDE procedures (WB-5112, WC-5112)	10 yr after superseded or invalidated
4. Personnel qualification records (WB-5520, WC-5520 and WB-4322, WC-4322)	3 yr after superseded or invalidated
5. Purchase orders	10 yr after superseded or invalidated
6. Final radiographs not covered in Table WA-4134.17-1	10 yr after completion
7. Calibration records (NCA-4134.12)	Until recalibrated

GENERAL NOTE: Nonconformance reports that affect those records listed and are not incorporated into the record shall be retained for the retention period applicable to the record the nonconformance report affects.

These tables show the exact document set that is required to be available for each nuclear facility. The nuclear industry's requirements for records appear far more normalized (i.e., have an effective minimum uniform dataset that everyone should meet) than the requirements of the pipeline industry. There are other examples from BPVC-III-D3 that show a higher level of normalization from the nuclear industry, such as specifications for nameplates, which are required to match a specified form-type for "each containment or part to which a Certification Mark is applied."

There are also examples of a higher standard of normalized documentation in the ASME BPVC-XI "Rules for In-service Inspection of Nuclear Power Plant Components." This standard provides standard forms for repair/replacement activities that include bill of material tables, which are required to be completed by the individuals that carry out the repair activities. The following figure, Figure 1-3, shows a direct capture of a portion of Report of Contracted Repair/Replacement Activity.

Figure 1-3: Partial Standard Repair/Replacement Form from BPVC-XI

5.	**Items Affected by the Contracted Repair/Replacement Activities**			
	Description of Item	Item Identification No. Assigned by Owner	Name of Manufacturer	Manufacturer's Model/Serial No.
(a)				
(b)	(5)	(6)	(7)	(8)
(c)				
(d)				
(e)				
(f)				
(g)				
(h)				
(i)				
(j)				

6.	**Items Installed During Contracted Repair/Replacement Activities**							
Identification			Construction Code for Fabrication of Installed Item					
Description of Item installed	Name of Manufacturer	Manufacturer's Model/Serial No. and Unique Traceability No.	Const Code/ Sect/Div.	Edition/ Addenda	Code Cases	Code Class	Installed into (Line No. from Section 5)	
(10)	(11)	(12)	(13)	(14)	(14)	(14)	(15)	

It may not be practical to expect a single list of document types to cover all required information about all types of systems, materials, and components that could populate a pipeline system during its lifetime. It also may not be practical to require a specific "form-type" of nameplate be attached to all pipe components, or a standard form to fit all pipeline repair/replacement activities, provided how much flexibility is required to handle the unique challenges within the pipeline industry. It does make sense, however, for the pipeline industry to move toward better normalization with improved and specific minimum data requirements and formats.

The nuclear standard BPVC-XI specifically requires a "records index" in the Quality Assurance Records section. The requirement states that "The records shall be indexed. The records and the indices thereto shall be accessible to the Owner, Owner's designee, and Authorized Nuclear Inspector." This type of accessibility allows for more effective decision making because the Decision Maker has access to the record itself and not just a transcribed version, like an alignment sheet. When a decision (i.e., fitness-for-service analysis or maximum allowable operating pressure (MAOP) determination) is made based on original construction documents, the Decision Maker can observe first-hand any items within the documents that could impact the specific decision, like the type of test that was performed or instrument used. In contrast, when the information is transcribed onto an alignment sheet, the ability to scrutinize the source is taken away from the Decision Maker.

Computer networking and software developments of the past one to two decades have provided the pipeline industry with an opportunity to change the philosophy of what is feasible from an accessibility standpoint for records. Twenty years ago, it wouldn't have been possible to provide source records digitally and instantaneously throughout an organization. In the past few years, pipeline operators have used the records-review initiative, following the San Bruno failure, to index their records thoroughly with document management systems and are using GIS to interface with records for specific pipeline component(s). This document management index provides them with an easier and more efficient document review, and the GIS interface provides the operator and the operator's designees (i.e., the Decision Makers) more efficient on-going access to the pipeline system data and records.

1.3 Recommendations

Practices from industries, such as the automotive, airline/aerospace, and food processing industries, were also reviewed, but the pipeline industry's unique set of characteristics made direct comparisons challenging. For example, the airline industry is highly critical to safety and is materials dependent, but all of the components of an airplane can be removed and/or inspected from a single hanger and without the need to excavate. This makes the data and recordkeeping needs fundamentally different from pipelines, which are hidden from view across a widespread geographic area and must be excavated to be inspected.

Lessons learned from studying international pipeline standards in Section 1.1, Background of Data and Recordkeeping in the Pipeline Industry, and from other industries reviewed in Section 1.2 have pointed to the following improvements that can potentially be made in the near term and longer term, including:

1. Normalize the data and record types that should be maintained.
    a. The authors propose a legacy normalization process that utilizes a common reliability ranking system (or document hierarchy; See Section 5).
    b. The authors propose a go-forward normalization approach of common minimum data forms and report templates like those found in the ASME BPVC standards for the nuclear industry.
2. Formally dedicate resources specifically to recordkeeping roles and responsibilities.
3. Provide more accessibility of the data and records to company deputies.

These lessons are likely to be a continuous improvement effort, as opposed to a one-time fix, but the text in the remainder of this guideline is provided as guidance to help the industry take the next steps to address these lessons learned.

The ISO standard 15489-1 would suggest the actions taken to make records more "useable" ensures "a user has the ability to access the necessary data within a reasonable time period." [24]

1.4 Recommendations for Revisions to ASME B31.4, B31.8, B31.8S, and Future Work

"Prior to pipeline regulations and modern evolutions of the ASME Code, there were no mandated levels of data quality, which suggests that the prevailing levels of quality in present data as a result of historical practices are no longer viewed by the public or regulators as adequate. Data quality requirements set forth by the Code today are overly general and are inadequate for the future. The issue then is what can ASME do to enable an operator to evaluate and assure data quality?" [25] ASME B31.4, B31.8, and B31.8S were reviewed for potential locations for revision based on the recommendations and lessons learned described in previous sections.

Current Revision Recommendations
Two locations in ASME-B31.8S show direct relevance, above others, to the lessons learned of the previous section. Paragraphs 4.3 Data Sources, and 4.4 Data Collection Review and Analysis each describe system-wide research efforts to collect and evaluate pipeline system data. Section 4.4 recommends "A plan for collecting, reviewing, and analyzing the data shall be created and in place from the conception of the data collection effort." ASME-B31.8S has provided specific guidance based on proven industry experience collecting, reviewing, and analyzing data (and rating it for reliability). It also provides specific guidance on

how the data should be stored and made accessible to company Decision Makers, how they should be cross-checked and updated, and how the team in charge of maintaining this pipeline system information database should be structured.

The contents in Sections 2 through 6 and Appendices A, B, and C of this guideline have more current data collection and recordkeeping guidance than what is currently in ASME B31.8S Sections 4.3 and 4.4. The plan consists of the following (by guideline section):

- Section 2 - An overview of the program with roles and responsibilities.
- Section 3 - A description to structure the pipeline information database to optimize access to source records.
- Section 4 - A process to determine likely boundaries for unknown values.
- Section 5 - A process to determine quality and reliability of data based on attributes of the source records.
- Section 6 - A process for on-going data collection and management of the records database.
- Appendices A, B, and C to support the main text.

The most relevant location for this guidance would be as a supplementary appendix to these sections in ASME B31.8S. Sections 4.3 and 4.4, which would remain in-full and a reference to the supplementary appendix could be added as the "recommended guidance to develop a plan for collecting, reviewing, and analyzing data."

Future Revision Recommendations

The guidance within this document is a product of the authors' and industry's experience over the last half-decade and was developed for MAOP verification. The focus of these efforts was typically on construction and maintenance records, which included pressure tests and material installation records. While the guidance in this document has been tested thoroughly for pipeline system materials and pressure test records, it has not been tested for some other types of data that ASME B31.8S Sections 4.3 and 4.4 require (e.g., SCADA, soil resistivity/moisture levels, one-calls). It is the authors' hope that the guidance and philosophies regarding accessibility and dedicated recordkeeping resources contained in this guideline could extend beyond materials and be adjusted to apply more effectively to other data types as future work.

During the review, there were many locations where records requirements were "overly general" as described above. A prime example follows, where the requirement is clearly adamant about the importance of detail, but provides no guidance on what details are needed or how they should be stored or tracked: "Records shall be made covering all leaks discovered and repairs made. All pipeline breaks shall be reported in detail." [26] Providing a form with minimum data requirements could provide operators with a valuable information gathering tool (particularly smaller operators who may not have the resources to call upon SMEs for each occurrence).

Meeting the spirit of the normalization standard set by the ASME-BPVC standards for the nuclear industry would require many specialized forms and report templates with minimum data requirements. Providing a recommendation on how to normalize each type of data needed for operations and maintenance and integrity and risk management practices in ASME B31.4, B31.8, and B31.8S is not possible within the scope of this guideline. Multiple subject matter experts will likely be needed to meet the normalization requirements consistent with the level of the ASME BPVC standards for the nuclear industry.

The remaining notable paragraphs from ASME B31.4, B31.8, and B31.8S that are candidates for normalization were itemized into tables shown in Appendix D.

2 PROGRAM OVERVIEW AND TYPICAL ROLES AND RESPONSIBILITIES

An operator who is interested in upgrading or maintaining a modern data and records management system as described in this guideline should include personnel and resources to fill the roles and responsibilities indicated below. These roles and responsibilities are referenced throughout the processes described within this guideline. The following roles[3] will need to be delegated to the appropriate personnel upon starting this process:

Table 2-1: Roles and Responsibilities

Role	Responsibility	Examples Title
Database Owner	Must be responsible for managing and maintaining the data within the Pipeline Database for the relevant section of pipe.	• Pipeline Integrity Manager • Pipeline Integrity Engineer • Risk Engineer
Records Librarian	Collection, storage, organization, and manipulation of data and/or documents (a data mining expert). Must be familiar with the hard copy records and "soft" data that is (or will be) used to populate the Pipeline Database. Must provide technical support to the Research teams about source records, their locations and formats.	• Pipeline Integrity Engineer • Risk Engineer • Administrator • GIS specialist
Subject Matter Expert (SME)	Provides extensive knowledge about the relevant integrity practices, pipeline construction and procurement, materials, company knowledge, etc. Typically provides technical support about this knowledge to others in the company.	• Pipeline Integrity Engineer • Procurement personnel
Records Engineer	Oversees teams of Researchers and Data Miners and performs quality assurance (including process controls and design) and quality control on their work.	• Pipeline Integrity Engineer • Risk Engineer
Researcher/ Data Miner	Conducts the necessary research into the background information related to records gaps or document reliability. Must be familiar with industry standards, compliance and regulator communications, and/or company information and documentation. Reviews new records/information and compares them to the database. Carries out the MOC process listed within this guideline and is responsible for resolving data discrepancies.	• Pipeline Integrity Engineer • Documents and Records Management Team • GIS specialist
Decision Maker	Makes decisions that require the information from the Pipeline Database, which may include critical operations decisions (e.g., pressure reductions and fitness-for-service analysis).	• Company management • Pipeline Integrity Engineer • Risk engineers
Company Management	Ensures compliance of the pipeline system, which includes the records and data management. Responsible for ensuring others on this list have the resources necessary to carry out their roles and responsibilities.	• VP of Engineering • Compliance Manager • Pipeline Integrity Manager

[3] Depending on the size of the operator and the number of records to maintain, it may be appropriate for some individuals to fill multiple roles.

3 COMPANY PIPELINE INFORMATION DATABASE

It is recommended that each operator maintain a chief source of pipeline system information, so they can make auditable, repeatable, and trustworthy decisions such as those needed for fitness-for-service calculations and risk assessments. The authors of this guideline recommend a practice of maintaining digital source records directly linked to specific pipeline components within an operator's GIS. This recommendation is based on a philosophy of providing the operator's Decision Makers easier access to the source records.

This system might not fit every operator's system. There may be some operators that can provide their Decision Makers ready access to all source records in a way that can be as efficient as directly linking the source records within GIS. This could be the case for a very small operator where records are always immediately retrievable for the entire system (and where it may not be practical to invest in the required network infrastructure and software considering the benefit).

This section describes guidance on how to create a company pipeline information database with digital source records linked within GIS. The guidance below is drawn from the authors' and the industry's experience working with pipeline system materials records in the past half-decade to calculate design pressures as part of verifying MAOP. In many cases, the philosophies provided in this guidance could be extended beyond materials and applied to other data types as well (e.g., SCADA, soil resistivity/moisture levels). The Work Flow Process Diagram in Section 3.3 summarizes the process described.

If an operator elects not to maintain this type of database, there may be many parts of this section that will not apply. However, as long as the operator maintains a chief source of pipeline system information (e.g., master alignment sheets) most of the guidance in subsequent sections will apply. This operator must know that when the authors reference the "company pipeline information database." the guidance is applied to the "chief source of pipeline system information."

3.1 Locations and Formats of Pertinent Data

The first steps in preparing a database of digital source records are to identify the types of records that are expected to be included and imported in the records database and to identify the locations of these records during a "pertinent data search," so they can be organized for import systematically. A key step in this process is recognizing which records are, or are not, pertinent or relevant. This, and other key steps, should maximize the value (amount of information identified as pertinent) relative to the effort required to retrieve and review records.

When identifying records that are expected to be imported and included, an operator should target locations of records that are essential to material properties verification. The locations may be buildings that serve as maintenance hubs or offices that belonged to a primary project manager (or project management team) for construction projects. The locations may also include record storage locations (either vendor or company owned).

The search team should research how the records have been stored and transferred through the history of the organization. Some key events that impact record locations are mergers and acquisitions (both of companies and of assets), closing offices down, or moving branch offices or headquarters.

This search usually begins at a primary, or representative, document storage location. Consideration should be given to communication and coordination with local personnel prior to arriving onsite. In many cases, the search team will be working with and removing documents that are critical to onsite personnel and/or local operations. Gaining cooperation from the onsite personnel could be critical to having a successful project.

Once the team arrives at a records storage location, the first priority is to become familiar with the onsite file organization. The onsite file organization can help discern between large groups of pertinent and non-pertinent information. For example, areas of the onsite file organization designated for "office maintenance files" are non-pertinent, and the team can save large amounts of time by avoiding further review. The effectiveness of this strategy is largely dependent upon the efficiency and organization of the onsite file structure.

This initial search should be a broad search and thought of as including or precluding data from being imported to the records database on a box-by-box, shelf-by-shelf, or bin-by-bin level or larger. It is important to note that this philosophy will also apply to softer media locations, such as compact disk files or microfilm and microfiche. If encountered, the team should endeavor to group records from these other media types and include or exclude them, when practical, in the same manner as the box, shelf, or bin level.

A search team should consist of personnel familiar with documentation for pipeline construction and maintenance. The team should be led by a Records Engineer with data mining experience, who can identify pertinent data types for the search. An ideal candidate for the team leader is an engineer that has led records verification projects, or an engineer with material experience as a project lead in pipeline construction and maintenance projects (or specifically the types of data that is being targeted by the search).

Project numbers can be the primary tool for linking (i.e., establishing traceability with) source record documents, such as purchasing records, hydrostatic test records, mill test records, and construction files to specific components or locations within a pipeline system. If the company has tracked information related to project numbers, this information could be useful during the pertinent data search. The utility of a list for this purpose will depend on the type of data that was tracked along with each project number. For example, if the operator tracked project type, or perhaps a short description with each project number, this information could be used to target projects that are more pertinent to the data needed or rule out non-pertinent data (i.e., a right-of-way land ownership dispute survey is not likely to produce any pertinent information related to pipeline system materials data).

Regardless of how helpful the project list is during the pertinent data search, it is good practice to begin a project number list during this stage in the project. It will also be used during later stages. A project number list should attempt to track at a minimum:

- Project number
- Project date (or year)
- District/area
- Pipeline name(s)/number(s), and
- Location/stationing.

Some operators have had success creating such a list from accounting records in addition to project management records. The project list can also be expanded as the document search proceeds. A robust project list will support locating pertinent data, document indexing (as metadata), and data mining of the records. The latter two of these items are discussed in subsequent sections.

The last type of check that can be used to discern pertinent from non-pertinent data is the quick peek check. This is the last filtering check if the onsite file organization and company project list haven't removed non-pertinent data at the box/shelf/bin level, then the team members will do a manual quick scan, or peek at the contents of each box, shelf, or bin to determine if the data is pertinent or non-pertinent. Each box, shelf, or bin that can't be concluded as non-pertinent will be imported to the records database.

3.2 Modernizing and Digitizing Data and Information

Once the proper records have been located, they should be digitally captured (i.e., scanned). Care must be taken to ensure that there is sufficient resolution to make quality copies. In the authors' experiences, 300 PPI[4] is an acceptable resolution for image capture. It provides adequate resolution, while not overburdening storage and network infrastructures with its file size requirements. It is important to have color scanning capabilities, but it is preferred to have the option to scan in color only when needed. This will allow for faster load times for black and white documents, whenever possible.

The authors consider the following guidelines [27] good practice for pipeline records digitization efforts:

- Digitize to the original size of the records.
- No cropping allowed – records should be complete.
- Back reflection of originals – Back all originals with bright white opaque paper.
- Scanning oversize items – oversize items may produce very large file sizes; legibility of small characters may need to be evaluated for resolution versus file size needs.[5]

Figure 3-1 is an example of a pressure test dead weight chart, which is scanned at 300 PPI and is color-coded to indicate temperature in one color and pressure in another. This image shows that the image quality is certainly sufficient for retrieval and review by Decision Makers. The color scan was critical to this image (to identify the red temperature trace with respect to the blue pressure trace). There is a portion of the chart that is truncated and although it appears there is no critical information missing, this should be avoided because future reviewers may wonder about the missing portion. Some digital document management systems will allow quality control comments to be stored as metadata with each record. These comments can be populated to assure future reviewers that there was no critical information missing, or no different information in the original. Populating this field can be time consuming, but it can provide useful quality control information if an operator would like a permanent digital file structure (as discussed below).

[4] PPI (pixels per inch) is often used interchangeably with DPI (dots per inch). PPI refers to the resolution of an electronic image, whereas DPI refers to the resolution of a printed image.

[5] Note that for pipeline records the size of the oversize file is often justified because the small print is often critical.

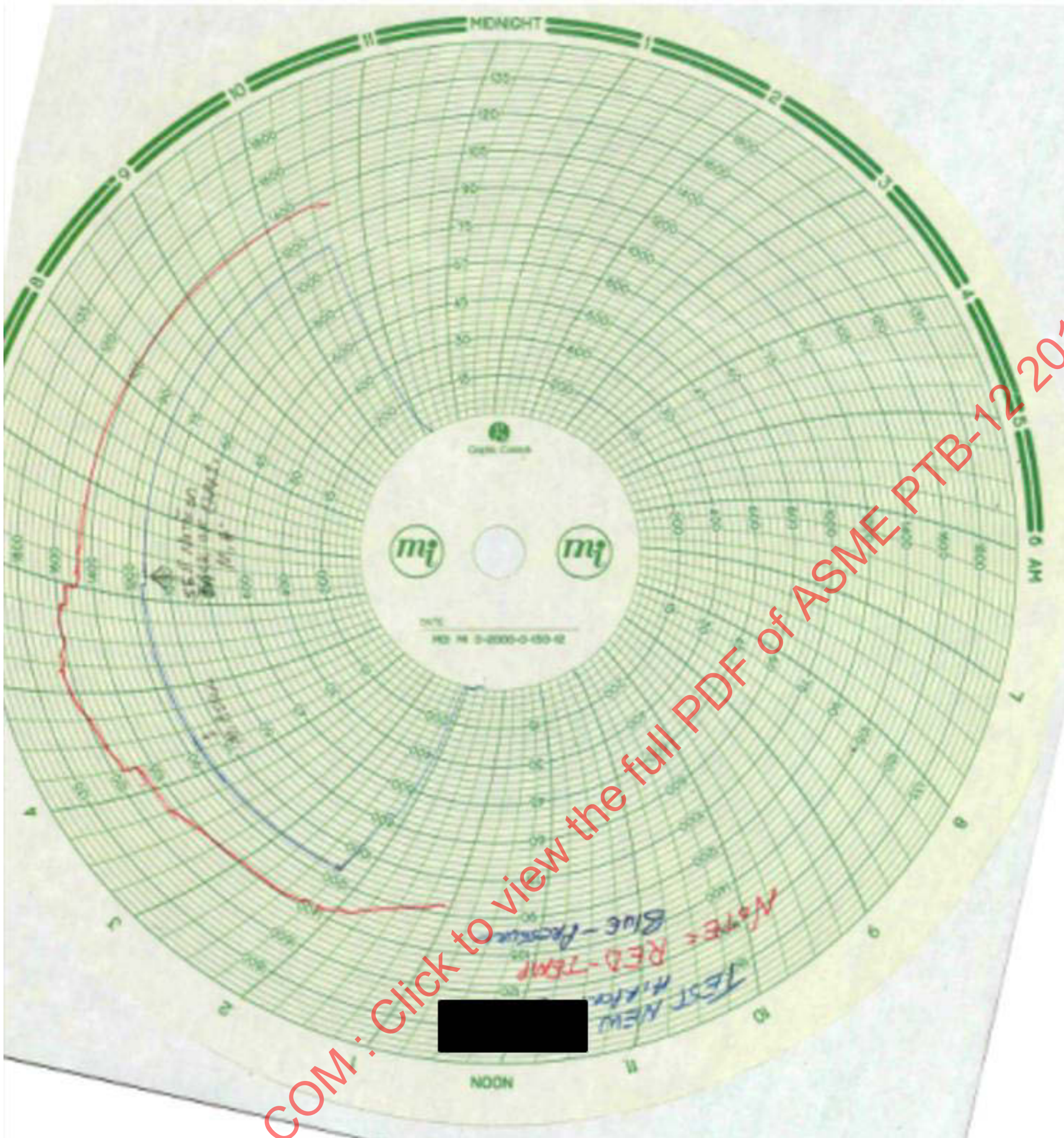


Figure 3-1: An Example of a Pressure Test Dead Weight Chart.

3.2.1 Document Indexing with Metadata

Digital document management systems, such as SAP®, OnBase™, and SharePoint™, are commercial options that have been successfully used to organize documents by document indexing with metadata in large-scale industry projects. Each has its own strengths and limitations. It is recommended that operators select a digital document management solution that is most compatible with their organization (or that they find a digital document management solution within their organization that is suitable to the requirements of this section).

The following parameters have been used by operators to select digital document management solutions:

- Security features
- Data backup features

- Customization capabilities for user-access (e.g., can it restrict certain documents, or groups of documents from certain users? Can it allow some users read-only access?)
- Ability to integrate with other software packages
- Ability to integrate across the operator's organization (i.e., is this a product that multiple departments will use?)
- Platform sustainability (is it likely to be supported by the publisher ten years from now?)
- Ease-of-use
- Cost (including on-going subscription fees)
- Data limitations
- Quality and capabilities of service providers associated with the platform
- Options for remote access (e.g., local network, intranet, or internet)

The records system described in this guidance report "requires a set of metadata and unique naming convention for each scanned document. The metadata and naming convention system should be designed by a (Records Engineer)." An effective indexing system should be consistent, understandable, and intuitive for Researchers and Decision Makers to target and retrieve documents. The metadata should allow the digital file structure to be searched and grouped dynamically. The metadata should include:

1. Unique record identifier (Document Identification)
2. Pipeline name or number (whichever is more likely to sustain over time)
3. Project number
4. Document type
5. Date created
6. Document author
7. Milepost or station number
8. Operator name
9. Hard copy location
10. Equipment or Joint Identification, if applicable
11. Version Number / current.

The first five indexing fields are the most important to an effort to systematically review pipeline system material records for design pressure data to contribute to MAOP verification. It is highly important to understand the chronology of the projects that occurred on each pipeline in order to understand the current configuration of pipeline materials (and replaced pipeline materials). The first five indexing fields allow a Researcher to understand the order of the projects and research where each project occurred within each pipeline.

Once a unique record identifier is assigned to a document, some operators have preferred to place the unique record identifier directly onto the hard copy document (i.e., with a barcode). This has provided them assurance that the hard copy will always be linked to the digital scan.

Other operators have felt that the bar code covers a portion of the document impacting its integrity. These operators have preferred to maintain the link between the hard copy and digital scan through metadata, by diligently tracking the hard copy location field, or through controlled version numbers.

As metadata and document naming conventions become more commonplace in the pipeline industry, there may be off-the-shelf options that are sufficient for some operator's needs. It would be advisable to conduct a short industry search to see if there are off-the-shelf metadata and naming convention options that an operator could incorporate. Most operators will have a sub-set of company specific naming conventions that will need to be taken into consideration (for example some operators may use "Valve Section" as a critical part of their stationing, while others may require a date to properly apply a "Version Number"). While off-the-shelf solutions can provide efficiencies, consideration should be given to implementation for a specific operator.

3.2.2 Digitization Quality Assurance

Large scale digitization efforts can be completed with the assistance of a scanning vendor. The scanning vendor will have procedures to carefully deconstruct the project boxes, folders, and binders to digitize the documents within. The indexing of metadata will be completed primarily by the scanning vendor (with some training on identifying the metadata within the pipeline records from the operator and/or Records Engineer).

The scanning vendor typically performs pre-digitization quality control checks. There can also be additional special handling requirements when analyzing old pipeline records. Some paperwork may not have been archived properly and may have been exposed to mold or degradation. Some paperwork may have faded over time, which will make a poor scan.

A scanning Quality Control (QC) process could check the following items to the original hard copy to ensure a quality scan process:

- Is the image accurate?
- Has the entire image been captured?
- Has it been correctly oriented?
- Is it centered?
- Is the order of pages correct?
- Are there contrasts or uneven tones?
- Is all writing legible?
- Is the color image true to the original (and how close does it need to be)?
- Has it been indexed correctly?
- What is the acceptable error rate for both accuracy and quality?

Some scanning vendors offer double-blind metadata indexing to double check their own work to ensure a very high level of quality. If the QC check is not required to be 100 percent, each operator must select its own acceptability levels for the QC check.

3.2.3 Disposition – Disposal of Data

Pipeline operators may be tempted to destroy original documentation once it has been digitized. Holding on to the original document could hamper document control, or it may be costly to manage. However, the original documentation could prove useful if future potential purchasers of assets have conflicting digital document policies, potential digital format conflicts, or if they wish to verify the digital document quality. It is conceivable that a potential sale would not occur if the asset information could not be independently verified.

There should be a plan in place for disposal of records, including who should make the decision, and justification standards. At the minimum, the plan should include a QC check of 100 percent of the digital copies with respect to the hard copy for the items mentioned in the section above, a plan to ensure the integrity of the digital copies with back-up storage location, and a plan to ensure the format of the digital records is stable from obsolescence.

3.3 Structuring Data and Data Sources to Maximize Access

Structuring data and data sources to maximize access is one of the most important steps an operator can take with its source records. This step can unlock the true potential and utility of a robust records library by putting the information in the Decision Makers' hands at the time when it is needed the most.

A Decision Maker can use GIS software to navigate a map interface to a particular section of pipe or a particular fitting and instantly retrieve attribute data from the pipeline system database (e.g., PODS, APDM, UPDM), which is a core function of GIS.

A primary recommendation of this guideline is to increase access to the source records that verify the pipeline system information (e.g., original construction records). Some operators are currently endeavoring to meet this recommendation by including "direct links" within the GIS interface to each verifying source record for the information in the pipeline system database. That is, once a Decision Maker navigates to a section of pipe and retrieves the material properties, he or she will also be able to instantly click and view the source records that verify each material property. The authors of this guideline have helped develop successful applications of this structure for pipeline system materials and pressure test information, including:

- Outside diameter
- Wall thickness
- Pipe Grade
- Seam type
- American National Standards Institute (ANSI) rating (if applicable)
- Test pressure
- Test date
- Test duration
- Test company
- Test medium
- Test employee/company.

This structure has also been applied for other information that appears in pipeline system data models and can be obtained from construction and maintenance files, such as coating type, girth weld type, and year of manufacture. It may be possible to use this type of structure (i.e., GIS interface with a link to other source data sets) for most, if not all, data sets that appear in the pipeline database model.

3.3.1 Data Mining and Gap Filling to Establish "Direct Links" to Source Records

The source records must be linked to the pipeline system attribute data by performing systematic records research (a.k.a. "data mining"). This can be very labor intensive, so it is recommended that a team of Data Miners conduct this step with oversight from a Records Engineer.

The data mining efforts can benefit from the digital document database in several different ways once it has been created. The digital document database can serve to centralize documents, whose hard copies were previously spread across several document locations, into a single data mining location.

The document index "metadata" that was recommended in Section 3.2.1 (and assigned by a scanning vendor) will also assist during data mining. The Data Miners should use the metadata fields to filter through the records and target the most relevant documents to research.

For example, the Data Miners could sort to a particular pipeline number to review all of the project numbers that have been indexed with that pipeline. The Data Miner could also review the dates indexed for each of the project numbers, which gives an idea for the chronology of the projects. Organizing the chronology of the projects within each pipeline is an important step in the data mining process. Many Data Miners have found that researching the projects in their chronological order provides a logical reconstruction of the pipeline materials verifying source records.

The project list that was mentioned in Section 3.1 could also be used during the data mining process. The project list can be used to help organize and perform QC checks on the chronology of the projects within each pipeline.

Section 5 is a discussion and methodology to create a data reliability index based on source record quality. This methodology is based around creating a document hierarchy or document ranking system. The document hierarchy or document ranking system plays a critical role in the data mining process, because it guides the Data Miners to the document types that could provide the most reliable data (i.e., the "high value targets"). It also provides a reliability metric for the information retrieved from source records, which is based upon the quality and characteristics of the source record itself. In general, if an operator is dedicating resources to perform a large-scale data mining effort, a document hierarchy will also be used to create a more efficient data mining process and to assign a reliability ranking to the data retrieved.[6]

Combining a document hierarchy with a document management system that indexes "document type" metadata can be a substantial time saving measure if the metadata "document types" have been indexed reliably to show the same types of document types indicated by an effectively developed document hierarchy. More details on data mining with a document hierarchy are presented in Appendix B.

The final goal of the data mining effort is to store the direct links between the attribute data and the verifying source records within the GIS system itself. During the data mining process, it may be easier to track results in a Microsoft Excel™ spreadsheet. This will allow the results to be adjusted and QC checked easily by the overseeing Records Engineer, when necessary. Once the results are complete and QC checked, they can be loaded into GIS.

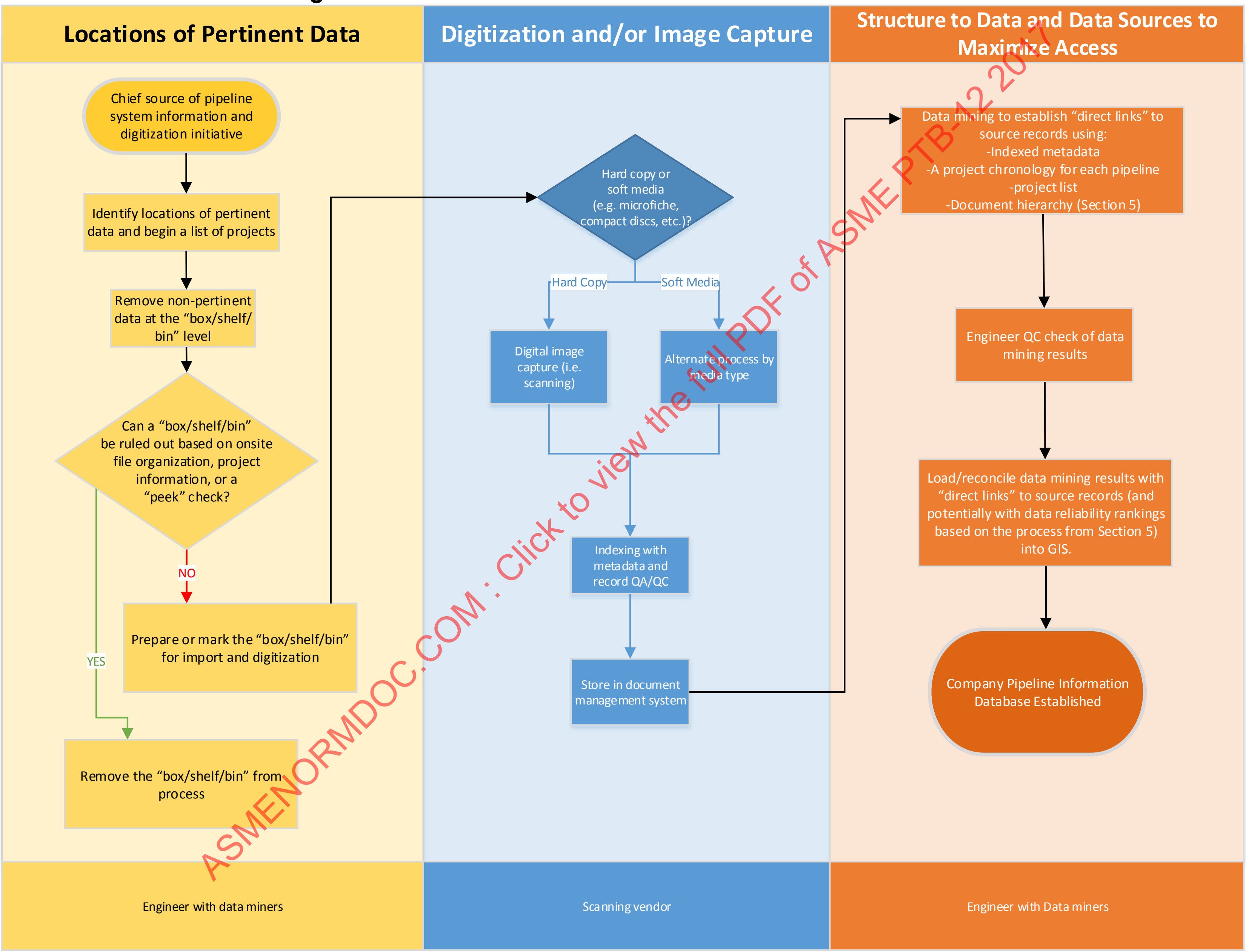
If the GIS system is populated with data, it may need to be reconciled with the data mining results. Depending on the level of confidence in the legacy GIS data, the data mining team may elect to incorporate the GIS data earlier in the data mining process and reconcile it along with the records review. If there is very little, or no, confidence in the GIS data, the team may elect to replace it entirely with the results of the data mining efforts.

Some GIS systems are directly compatible with digital document management systems (for example, OnBase™ publishes a specific module for ESRI), while other GIS systems (e.g., Uptime™) will accept URL address links to documents. Either method will allow documents to be directly linked. This type of records management (like a GIS system itself) must be maintained by the Database Owner or the Records Librarian. Section 6.1 discusses the resourcing requirements for on-going records maintenance in more detail.

If the data mining efforts fail to fill all gaps, it is likely that default values (see Section 4.3) will be used to supplement the missing information. Use of default values must be specifically tracked. If new information becomes available, the default values may need to be systematically changed, or in some cases replaced with data from source records.

[6] Pipeline operators in the U.S. have been using this approach to satisfy PHMSA's requirement that pipeline system design data that contributes to verifying MAOP be "traceable, verifiable, and complete."

3.4 Work Flow Process Diagram



4 ESTABLISHING LIKELY BOUNDARIES FOR UNKNOWN VALUES

Establishing likely boundaries for unknown values will generally follow an operator's decision to initiate efforts to address gaps within the company pipeline information database. This decision is often based on a substantial change within the company or industry such as:

- Merger or acquisition
- Large expansion of assets
- Change in regulations, or industry standard.

This decision may also be the result of an identified deficiency, for example:

- Regulatory audit finding
- Self-audit finding, or self-identified gap.

In the worst case, this decision may be the result of a failure. The process to establish likely boundaries is described in Sections 4.1 through 4.3 and shown schematically by the flowchart in Section 4.4.

4.1 Database Retrieval

The Database Owner should be responsible for data gathering with potential support or input from company SMEs. The data targeted for gathering and the specific SME support will depend on the nature of the gap identified. Some gaps may relate to a short replacement of one specific area on a pipeline, which may only require input from one engineer or project manager. Other gaps may be related to a system wide deficiency in verified pipeline materials information, which might require input from the pipeline integrity manager or a pipeline materials expert.

The Database Owner will capture a broad view of the data available by exporting all relevant known information from the pipeline database (i.e., if gaps have been identified with pipe material properties, then the initial step is to export and review the material properties that are present in the database). If the Database Owner suspects he/she is missing relevant information within the database, an SME may be able to help identify the location(s) of the relevant information. If information is still missing after the SMEs have been informed, it should be assumed as a gap.

The Database Owner, SMEs, or both will examine the exported data for missing and unknown values that would indicate the information needs of the company haven't been met. The following questions may help determine if the exported data meets the information needs:

- Does the data contain gaps that substantiate the concerns that initiated the review?
- Does the data contain gaps that prevent the operator from determining required operating parameters, such as maximum allowable operating pressure?
- Does the data contain gaps that prevent the operator from performing accurate risk assessments on the pipeline system?
- Does the data contain gaps that prevent the operator from performing fitness-for-service assessments (e.g., ASME B31G burst pressure evaluation)?
- Does the data contain gaps that prevent the operator from effectively implementing their integrity management program or other necessary programs and procedures?

If the Database Owner and SME agree that the exported data does not meet the operator's information needs, then conservative boundaries of likely values must be identified to allow critical calculations and evaluations to be performed when data gaps exist. The Database Owner will proceed to research the appropriate range(s) for values of the unknown parameter(s) in the next section of this guideline.

If the Database Owner and SMEs agree that the exported data meets the information needs of the company, then there are no gaps in the exported data and the Database Owner will proceed to the quality and reliability review covered in Section 5.

4.2 Research Gathering

This section is the research phase and is expected to take the majority of the time when bounding likely values. The personnel required for this portion are the Database Owner and Researcher(s) with input from an SME.

The purpose of the research is to gather information that allows the team to assign conservative, realistic ranges of values for missing parameters. For example, if an operator identifies a gap in wall-thicknesses for parts of the pipeline system, this could trigger a review of historical manufacturing practices (American Petroleum Institute (API) API 5L and API 5LX) to identify the range of wall thicknesses provided by industry standards. It may also trigger a review of historical purchasing records to evaluate the range of wall thicknesses that have been purchased (and documented). The company may also survey field engineers and technicians about the range of wall thicknesses within the system according to their knowledge.

Note that the SME provides input and support in multiple ways in this section. It may be appropriate for a company's SME to provide the Researchers with some background knowledge on the research topics and provide some direction prior to conducting research (i.e., which sources should be reviewed first, help estimating research efforts, etc.). In some cases, the SME may be able to use the exported data discussed in the previous section of this guideline to help narrow the scope of research to make the efforts more efficient.

The SME may provide guidance for the Researchers during their review, including insight and confirming the research results (Was the research material gathered from the right locations? Was the information from a credible source?).

The research should include reviews of source material from the following four categories:[7]

1) **Industry standards** should be reviewed to establish standard practices and known physical limitations related to the gap data. Typical industry standards may be published by the following sources: ASME, API, Pipeline Research Council International (PRCI), and NACE.
2) A compliance review should be performed, which should include **regulatory requirements and communications**. This review will ensure that any relevant regulatory concerns are addressed and that the boundaries placed on likely values do not create a compliance issue. Sources for regulatory requirements include CSA Z-662, the CFR, and applicable state regulations. Sources that should be reviewed for regulatory compliance guidance include PHMSA Advisory Bulletins, frequently asked questions, and PHMSA's integrity management inspection protocols.
3) **Company information** should be reviewed to understand the range of documented values within company records. Pertinent records may include: As-built drawings, Purchasing Catalogs, Process and Procedure Documents, External and Internal Studies, and Operating Records (e.g., of current or past MAOP/MOP settings).

[7] This step is highly critical for each gap type. For example, stop-gap values determined from a paper study described in this section (including Bayesian Network approaches) may not meet minimum industry safety requirements or U.S. regulatory requirements for "traceable, verifiable, and complete" until fully verified with field measurements.

4) **Subject Matter Experts** with extensive knowledge about the pipeline system and/or subject matter should be consulted to confirm the ranges of values are realistic and applicable to the pipeline system. In some cases, the SMEs can also provide ranges of values based on their knowledge and experience. SMEs that may help include, but are not limited to, pipeline and risk engineers, purchasing, GIS, consultants, pipeliner/technician, etc.

Examples are provided above for each of these categories, however there may be other relevant organizations and sources that have relevant information. The utility of the research performed on each category will be depend on the nature of data gap, reinforcing the importance of consulting an SME prior to starting the research to provide the Researchers with some direction and background knowledge.

Once ranges of values for missing parameters are identified through the research described above, the Database Owner will use the research results to determine the boundaries to use in the company database in the next step.

4.3 Determine Boundaries

The personnel required to determine and approve the boundary limits are the Database Owner, SMEs, and potential Decision Makers including Company Management.

A list of all possible values (or ranges of values) will be established for the unknown data points. Unrealistic values will be ruled-out, whenever possible, and justified based on the research that was gathered and input from SMEs. The remaining values will be the possible "default values" that will be used with conservatism in the case of a null or gap in the company database.

It is advised to determine at least two individual values for each potential gap or null in the company database. For example, if the operator has identified a gap in wall-thickness values in parts of their pipeline system, he may elect to establish default wall thickness values for each pipe size: One that represents the "most likely case" value and the other that represents the "worst probable case" value.

The "most likely case" may be the wall-thickness that is most common for a particular pipe size, while the "worst probable case" may be the thinnest wall observed in API 5L and API 5LX for each pipe size. By selecting the "most likely case" and "worst probable case" values, the team can evaluate the impact of the uncertainty against conservatism for the default value(s). In some cases, conservatism may dictate that the operator should only use the "worst probable case" and not justify consideration for a less conservative "most likely case." In other cases, for example in some risk models, it may be more appropriate to utilize "most likely cases" to ensure the "worst probable cases" don't skew the results away from true issues of concern.

The "most likely case" and "worst probable case" values can be determined by one of the following methods:

- Review the range of possible "default values" with an SME who is knowledgeable about the subject matter to select what the "most likely case" and the "worst probable case" value should be.
- Using information from other similar pipelines or segments of the same line to determine the most common ("most likely case") value and maximum/minimum ("worst probable case") value across the pipeline system. The values that determine what a "similar" pipeline/segment is will depend on the nature of the data gap.
- Create a Bayesian Network Model to determine the uncertainty between the different values and then use those results to select the average ("most likely case") value and maximum/minimum ("worst probable case") value. Bayesian networks are powerful tools for filling gaps in pipeline databases, particularly when used for risk model applications. Bayesian networks calculate the

probability of different states (values) of unknown parameters. This method is most effective when there are multiple known variables that have an implied or direct influence on the unknown (gap) variable. Bayesian analyses have been used for applications in the pipeline industry during risk determination, root cause analysis, or maintenance prediction. More details are given in Appendix A.

Regardless of the method used, it is imperative to document the decision making process. The reasoning for choosing and ruling out values should be tracked, so that if additional information becomes available, the data can be revised.

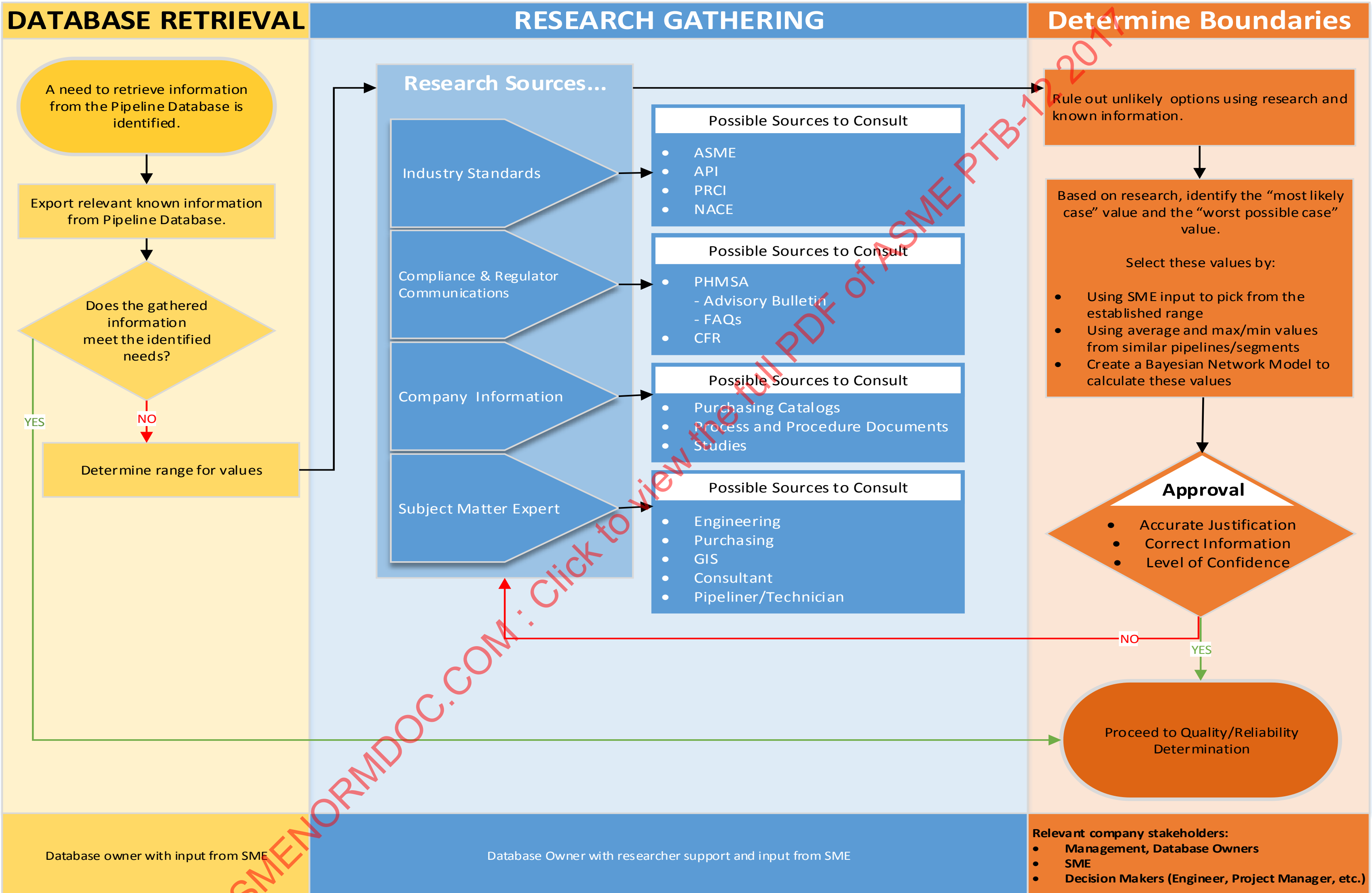
As discussed in Section 3.3.1, use of default values must be specifically tracked. If new information becomes available, the default values may need to be systematically changed, or some cases replaced with data from source records.

The options presented above increase in detail and comprehensiveness from top to bottom, so the team should select the appropriate method for each data type that is reviewed. Once the "most likely case" and the "worst probable case" values are selected, they should be reviewed for approval.

- If Approved: The previously unknown values have been assigned a "most likely case" and a "worst possible case." The team will create a report to document the research and decision making criteria used and any tables or lists that show the most likely and worst case values used. It will then proceed to the quality and reliability review covered in Section 5[8]
- If Rejected: One/Both of the values selected do not meet one of the decision criteria. The Decision Makers and SME will provide feedback to the Database Owner and Research team and they will need to return to the research stage in order to obtain additional information to address the feedback received and complete the re-attempt to determine boundaries for the data gap.

[8] Note that in the authors' experience operators have used this approach to fill data gaps until verification measurements (i.e. direct examinations and material tests) can be completed. No operators have interpreted this method to satisfy regulatory "traceable, verifiable, and complete" requirements (to the authors' knowledge), so it is highly recommended these entries are tracked as "assumed worst probable case," or "assumed most likely case."

4.4 Work Flow Process Diagram



5 ESTABLISHING DATA QUALITY AND RELIABILITY

Establishing data quality and reliability will generally follow a company decision or requirement to determine the quality of a source record(s). The decision to determine the quality and reliability of source record(s) is often based on a substantial change within the company or industry such as:

- Merger or acquisition
- Large expansion of assets
- Discovery of New or Different Industry/Recommended Practices.

This decision may also be the result of an identified deficiency, for example:

- Regulatory audit finding
- Self-audit finding
- Discovery of Conflicting or Errant Records.

In the worst case, this decision may be the result of a failure. The process to establish data quality and reliability is described in Sections 5.1 through 5.3 and shown schematically by the flowchart in Section 5.44.4.

5.1 Document Type Review

The Database Owner along with the Records Librarian should compile a list of all types of source records that contain information or data that populates the pipeline information database (e.g., construction project records that show pipe grades, wall thickness, test pressures, etc.). The Database Owner and/or Records Librarian should also consult SMEs (e.g., engineers who prepared construction project files for audit or permanent storage) to provide insight about the types of source records the company has used to document pipeline system information. Examples of source records include construction reports and notes, final drawings, design pressure form, mill certificates, materials testing reports, maintenance reports, and materials invoices. It is also recommended to search through construction project files from different geographic locations, legacy companies (if acquisitions took place), and time periods as a way to capture a variety of source record types. Once a complete list of source record types is thought to be compiled, the Database Owner and Records Librarian should meet with an SME to confirm the list.

There should be a pre-existing process to determine source records' quality and reliability. The Database Owner should examine company documentation to determine if the company already has an acceptable method to determine the quality and reliability of a source records.

The following questions may help determine if a quality and reliability determination process is acceptable.

(1) Does the existing process comply with the most recent practices and communications from regulatory entities?
(2) Does the existing process incorporate the source record types identified from the search described above?
(3) Does the SME agree with quality and reliability ranking based on his/her experience with the company's source records?
(4) Does the process establish traceable, verifiable, and complete documents?
(5) Is the process adequate to resolve the development mentioned at the beginning of this section, which triggered the need for metrics for data quality and reliability?

If the existing quality and reliability determination process is found unacceptable or does not exist, then a new process must be created (see next section). If the existing quality and reliability determination process is found to be acceptable, then it may be used.

5.2 Research Gathering

The research gathering phase of the quality and reliability metric determination process is a crucial stage in developing an acceptable process. The personnel required for this section are the Database Owner and Researchers with likely input from an SMEs.

The purpose of the research is to gather information that allows the team to create a compliant process that is suitable for the company. The team must achieve an understanding of which source record types (i.e., document types) should be categorized with high quality and reliability. The quality and reliability of source records should relate to the attributes of the records including the intent of the documents' creation. To achieve an understanding of which source record types should be categorized with high quality and reliability, these attributes must be well understood for all record types.

In many cases the types and formats of source records have significant differences for different periods in time or for different geographic areas. Depending on the company size and structure, it is possible that a certain source record type may require different level(s) of quality and reliability in different districts of the company. For example, a company geographic region or districts may have historically emphasized a 'Project Completion List' as the most important source record for recording pipe and fitting materials data for company records. But another district may have typically emphasized a "Completion Drawing" as the most important source record for permanently recording pipe and fitting materials data.

The quality and reliability of a source record type will be influenced by the person who was responsible for creating it and how long after the installation the document was created. A "Project Completion List" created by a project inspector immediately upon completion of a construction project is generally considered more reliable than a materials list that was transcribed years after a project and by someone with no direct knowledge of the construction.

Other causes of internal record quality variances include:

- Records organization/storage practices
- Manufacturer/Supplier documentation practices
- Completion markings (i.e., stamps, dates, and/or signatures).

These types of variance should be researched and taken into account when creating acceptable metrics to determine quality and reliability of source records.

The research should include a review of recommendations on records quality from regulatory bodies and other industry organizations like those that followed the San Bruno pipeline failure. PHMSA released two advisory bulletins and a letter to the American Gas Association (AGA) [15][16][31] describing clarifications of regulation to ensure a line is being operated at a verified MAOP/MOP.

The practice to verify MAOP/MOP is to confirm an operator possesses the information necessary to satisfy applicable regulations (e.g., CFR 192.619 for gas pipe in the U.S. and CFR 195.406 for liquid pipe in the U.S.) with data that can be verified by source records that meet regulatory requirements. Source records that verify MAOP/MOP may include hydrotest records and/or construction materials records that provide the parameters used to calculate the design pressure (CFR 192.105 for gas pipe in the U.S. and CFR 195.106 for liquid pipe in the U.S.). PHMSA stated that if an operator is using records to determine the MAOP/MOP of a line segment, then the referenced records must be "traceable, verifiable, and complete." PHMSA provided the following descriptions of the criteria for these traits:

- **Traceable** meaning the record can be "clearly linked to original information about a pipeline segment of facility." Examples given included: Pipe mill records, purchase requisition, as-built documentation.

- **Verifiable** meaning the record contains information that is "confirmed by other complementary, but separate, documentation." Examples given included: Contract specifications for a pressure test complemented by pressure charts or field logs.
- **Complete** meaning the record is "finalized as evidenced by a signature, date, or other appropriate marking." Examples given included: pressure testing record should identify a specific segment of pipe, who conducted the test, the duration of the test, the test medium, temperatures, accurate pressure readings, and elevation information as applicable.

PHMSA further clarified "verifiable" in a follow-up letter to AGA. PHMSA confirmed that "verifiable" source records can be a single quality, traceable and complete record or a record confirmed by other complementary, but separate, documents. The letter does not define criteria about PHMSA's interpretation of "quality," other than "a single quality document that is traceable and complete, as evidenced by appropriate markings, would be acceptable."

Criteria that operators have used to determine a quality record include, but are not limited to, who created it, when it was created, the intent of the document (i.e., document type), and the amount of information filled in (with respect to what was expected/required to be filled in), presence of a title block, presence of a document date, completion stamp, or a signature. The extent to which each of the above criteria determines a quality record generally varies by operator.

Before starting the research process the SMEs and Database Owner should delegate areas and provide the research team with a plan of action to minimize overlapping topics. The research should include, but are not limited to, reviews of the following:

(1) **Industry standards** should be reviewed to determine if there are standard practices or document quality trait requirements. It is recommended to look in some of the typical industry organizations such as ASME, API, and PRCI, but there may be other organization related to other industries that have some general guidance on data management practices and record quality verification.

(2) **Regulatory requirements and communications** should be reviewed to ensure the quality and reliability determination method to be created will not create a compliance issue. In addition to the sources mentioned above, PHMSA may have released additional Advisory Bulletins, frequently asked questions, or other communications that contain interpretations of record quality and verification. It is recommended to review the code of federal regulations and applicable state regulations as they may also provide compliance requirements for quality of source records.

(3) **Company information** should be reviewed to understand previous company decisions and trust in the different records source types. There may be other quality processes and practices utilized in other areas that could be applied to this process to ensure that accurate information is being used. As demonstrated in previous examples, different segments of the company may have different experiences with record type quality.

(4) **Subject Matter Experts** familiar with company quality processes as well as those with a substantial amount of experience reviewing and working with the different source record types should be consulted to provide their insights specific to the company and the other research gathered. The type of people could include pipeline and risk engineers, purchasing, GIS consultants, pipeliner/technicians, and/or documentation managers.

Once the quality and reliability criteria requirements and recommendations are identified through the research described above, the Database Owner and SMEs will work together to develop specific ranking criteria for source records in the next section.

5.3 Determine Quality and Reliability Metrics

The personnel required to determine the Quality and Reliability Metrics are the Database Owner, SME(s) (i.e., those familiar with the company's regions and/or history, and familiar with construction procedure development), and Decision Makers including company management.

The team should review the quality and reliability criteria requirements and recommendations from the research step (Section 5.2) to create a methodology to determine data quality and reliability. The team will decide on appropriate criteria to rank, rate, and grade the source records, which will determine the quality and reliability of information extracted from them and placed into the pipeline information database. The examples below present different types of quality and reliability ranking systems.

One potential quality ranking method ("Method 1") is a relative ranking system by source record type. This method requires the team to rank all source record types from most reliable ("1") to least reliable ("15"). This type of listing can be very easy to interpret, but it places nearly all emphasis on source record type (as opposed to other potential attributes). This is appropriate when the source record types are very similar within a pipeline system and have very little variation (i.e., if construction project files within a geographic region were created with a high degree of consistency). If this method is used, it is important to have an excellent understanding of which source record types have been the most trustworthy in the construction project files.

The ranking of source record types may need to be adjusted to account for the specific data being extracted. For example, a visual inspection document may provide reliable information about coating type and rank very high ("1"), but it provides less reliable information about yield strength and would rank low ("15"). Separate rankings may be needed for each data type to resolve this discrepancy. The ranking of source record types may also need to be adjusted to account for variability between time periods or geographic region (i.e., "project completion list" may have a "1" ranking for the North region, but the "as-built drawing" may carry a "1" ranking for the South region). This method may require a lot of deliberation between the Database Owner, SMEs, and Decision Makers to create a final list(s). Some advantages and disadvantages of "Method 1: relative ranking by source record type" are provided here:

Advantages:

- Easy to use during data mining because it requires very little interpretation.
- Results are highly repeatable.

Disadvantages:

- Has the potential to overemphasize source record type above other criteria.
- May require a lot of deliberation to adequately frame up list(s) to consider all relevant data types and legacy influences (i.e., time period, geographic region).

An abbreviated example of a document hierarchy using Method 1 is shown in Table 5-1**Error! Reference source not found.**.

Table 5-1: Method 1 Document Hierarchy Example. Relative Ranking by Source Record Type

Ranking	Source Record Type	Number of Required Documents
1	Materials Certification	1 required
2	Completion Drawing	1 required
…	…	…
15	Notes from Field	2 required

Method 1 is best suited when the source record types are all consistent with respect to the type and detail of information contained in each individual record (i.e., if the Materials Certification and the Completion Drawing consistently meet company records requirements, like "traceable, verifiable, and complete"). If this is true, then the benefit of creating a Method 1 process is that it is very straightforward to use once it is set up.

It is recommended that Data Miners review each individual document being evaluated to ensure documents are consistent with expectations and/or any standards set by the operator based on research findings (i.e. no missing signatures, etc.). This type of document hierarchy tends to be specialized to the segment of the company that created it. It is recommended to review source record type quality assumptions before use in a different company segment. If the intent is to create one method (as opposed to multiple ranking sheets), or if the company's source records are found to lack consistency, then the relative ranking by source type method may not be appropriate.

Another quality sorting method is to develop bins based on potential source record traits ("Method 2"). This method is more flexible than the discrete rankings shown above and less dependent on the source record type. Based on the research conducted, there are specific traits that have been assigned to evaluate the quality and reliability of any given source record. These traits should include, but may not be limited to the following:

- Time that the record was produced
  - Was it prior to project completion?
  - Was it long after project completion? or
  - Was it in the timeframe right after completion when information would be most reliable?

- Creator of the document
  - Is the creator of the document noted by name and title?
  - What was their relationship to the project?
    - Was it created by a clerical administrator without any design or materials experience?
    - Was it created by a project inspector who would have direct knowledge of the materials installed?
    - Was it created by a project engineer who would have direct knowledge about the project design process?

- Intent of the document
  - Was the document intended to be a rough preliminary costing tool for the project?
  - Was the document intended to be the permanent record of materials installed by the project?
- Formality of the document (related to intent)

  - Was the document created on a personal journal or notepad (i.e., intended for personal reuse)?
  - Was the document created on a company letterhead suitable for distribution?

- Completeness of the document
  - Is the document signed, dated, or stamped in any way?
  - Does the document have any fields where a signature, date, or stamp are clearly expected or required but are missing?

Table 5-2 shows an example hierarchy utilizing some of the traits above. Some advantages and disadvantages of "Method 2: Source record bins according to specific document attributes" are provided here:

Advantages:

- Can be applied to documents that have a large amount of variability (i.e., different source record types or variability in a single source record type).
- This is typically easier to develop (compared to the Method 1) document hierarchy.

Disadvantages:

- Requires the ranking of each document to be interpreted.
- May be less repeatable if two individuals identify attributes inconsistently.

Table 5-2: Method 2 Document Hierarchy Example. Source Record Bins According to Document Attributes

Bin 1A		Additional comment/example:
Common Examples:	Pressure Test Assessment Commissioning Pressure Test Chart	
Timeframe:	Created during the process or within 1 year	
Creator:	Someone on site	Crew foreman Inspector Project Manager
Completeness:	Must have all expected information	Clear location (Lat / Long) Project Manager Date Signature/Stamp

Bin 1B		Additional comment/example:
Common Examples:	Work Order As-built/Completion drawing	
Timeframe:	Created upon project completion	
Creator:	Someone with direct association to the project	Engineer Project Manager
Completeness:	Must have all expected information	Clear Description Date Signature/Stamp

Bin 2		Additional comment/example:
Common Examples:	Alignment Sheet data	
Timeframe:	Created within years after completion	
Creator:	Someone with direct association to the project	Engineer Project Manager
Completeness:	Needs only the critical expected information	Clear location (Lat / Long) Project Manager Date

For the sake of the example there are three bins shown. It may be necessary to create additional bins to adequately cover all ranking criteria. The bins above are shown as Bin 1A, Bin 1B, and Bin 2. Bin 1A represents high quality pressure test documents that meet the operator's "traceable, verifiable, and complete" document quality requirements for pressure test data without the need for any separate verifying records. Bin 1B represents high quality documents that meet the operator's "traceable, verifiable, and complete" document quality requirements for pipe material property data without any separate verifying records. Bin 2 represents documents that will only meet the operator's "traceable, verifiable, and complete" requirements for pipe material property data when accompanied by a separate complementary verifying source record (which may also be another Bin 2 record).

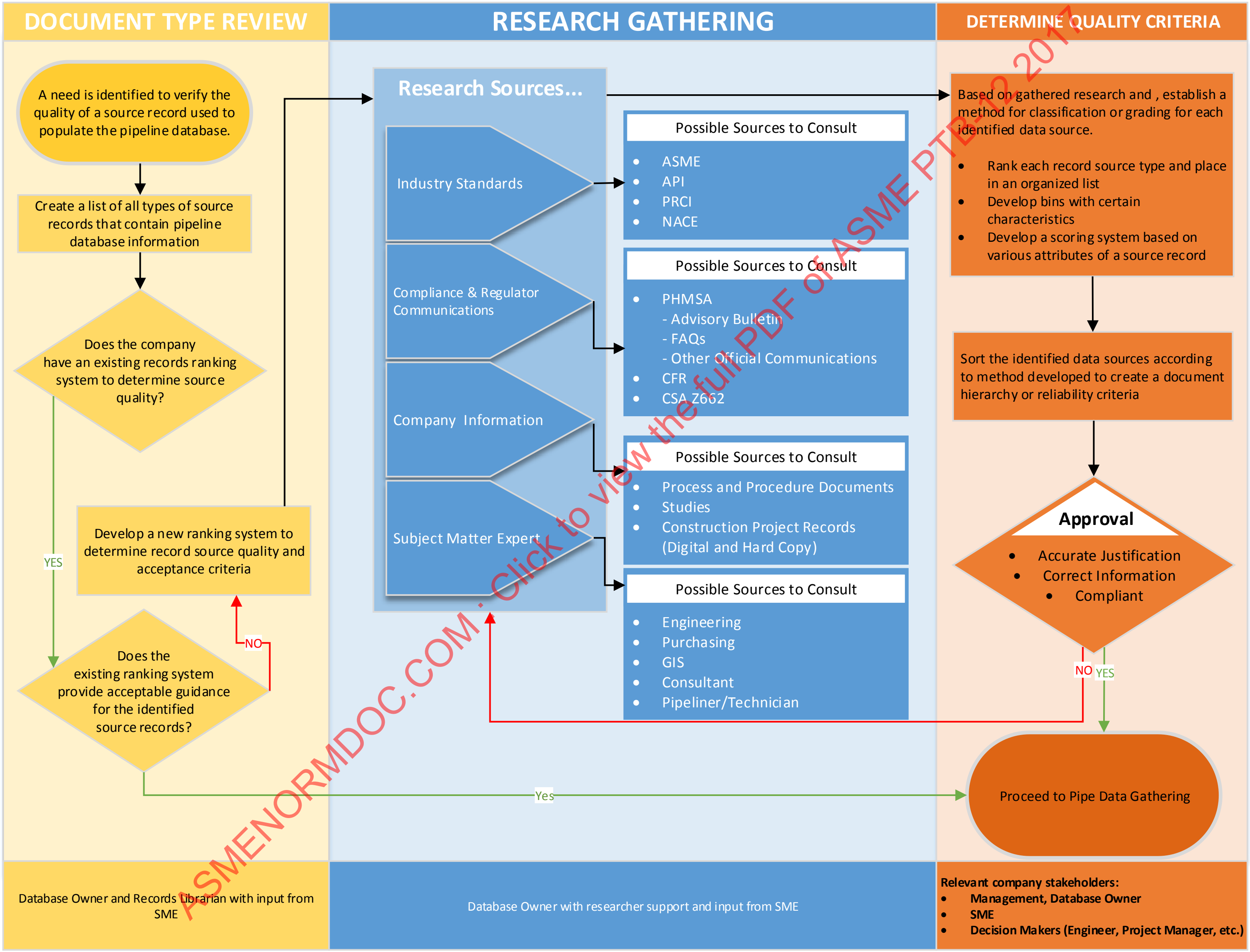
The benefit to this type of model is the quality checks intrinsically built into the task of matching a bin to a specific source record, however, due to this flexibility it can become more time consuming when attempting to verify a source record's quality and reliability if there is no perfect match to a specific bin. Operators have used both of the methods above to denote how many verifying source documents are required to meet company needs.

Typically, in either Method 1 or Method 2, operators denote a "high quality" document category, where one document is required to meet company reliability requirements (e.g., with "traceable, verifiable, and complete" source records). Operators will also denote requirements for a records category that must be verified by other records to meet minimum company reliability requirements. These two document categories are denoted "Bin 1 (A or B)" and "Bin 2" in the Method 2 example above.

After the process is created, it is highly recommended the team that developed the process should proceed to perform a test pilot to sort through source records and apply the methodology. During this task, the team should pay close attention to the ease, efficiency, and intuitiveness of the metric determination process developed. If the team identifies issues that hinder or prevent source records from being evaluated as intended, then process adjustments should be made. Once all necessary adjustments are made, the quality and reliability determination process is finalized and approved.

- If Approved: The quality and reliability determination process that has been developed will be put in use to verify the quality and reliability of source records for data in the pipeline information database.
- If Rejected: The quality and reliability determination process remains unacceptable. The Decision Makers, Database Owner, and SMEs will need to conduct more research and/or adjust the process for improvement.

5.4 Work Flow Process Diagram



6 ON-GOING DATA COLLECTION

Data collection, as it pertains to information about a pipeline system, has always occurred continually in the pipeline industry. Operators and their contractors are constantly recording observations and documenting measurements as part of field, survey, and testing activities. These new observations and measurements in effect create new source records with respect to the pipeline system database. An on-going data collection process is described in this section to maximize the utility of this data towards filling gaps, increasing confidence, and maintaining the database into the future. The process is shown by a flowchart in Section 6.5.

This process will require support from Data Miners to collect the incoming data under the oversight and coordination of the Database Owner and SMEs.

6.1 Sources of On-Going Data Collection

An operator must identify all practical sources of ongoing data collection (that can be integrated back to the pipeline information database). Since the number of potential sources of data can be large, this can be a challenge and will likely require work from Data Miners with input from SMEs (about the best data sources to fill specific database needs). It is recommended that the sources of ongoing data collection should include (but are likely not limited to):

- Direct examination reports that record observations made or measurements taken about attributes in the pipeline information database (e.g., observations on coating type, or wall thickness measurements).
- Maintenance reports that record new materials installed onto the pipe (e.g., a valve replacement report).
- ILI reports that document information along the pipe. The exact information depends on the ILI technology (e.g., wall thickness changes).
- Cut-out mechanical test results (e.g., Charpy toughness).
- As-built reports for new build and replacement projects.
- GIS centerline survey reports (could be part of ILI, or CIS surveys).

The sources of incoming information can be used by the Data Miners to fill data gaps or increase an operator's confidence in areas of the pipeline information database where the legacy source records provided low reliability.

In some cases, the information may have documented new materials and infrastructure that were installed, which will call for effective management of change for the new source records to replace the old source records. In other cases, measurements and observations are made about the existing materials and infrastructure, which can be used to confirm and increase confidence levels in the existing materials database. These scenarios are discussed in the following paragraphs.

6.2 Controlled Process for Management of Change

If an operator installs new materials or replaces materials, the operator must ensure that the new source records adhere to the established system for records management. The stakeholders for the project that generated new source records must be made aware of the established system for records management. There must be a solid understanding of how the records quality is rated.

Ideally, new projects should always produce records of the highest reliability. This will hold true if there is a solid understanding of the expectations for the new source records. For newly replaced or installed pipe, it is recommended that each operator prepare as-built reports that include the following specific parameters:

- Weld map(s) that reference a heat number for each joint of pipe installed.
- The weld map(s) must have sub-meter GPS coordinates for each end of a replacement segment, every angle point, or every half-mile (whichever is more frequent).
- An accompanying MTR(s) that includes each heat number that appears within the as-built.
- A dated signature from an onsite inspector specifically to confirm each joint was installed as drawn. The signature date should follow (or be coincident with) the date the joint was welded in place.

These parameters should help meet the spirit of normalizing (standardizing) as-built documentation of materials for installed pipe and set expectations between the construction team and the data mining team. Similar parameters are sufficient for elbow fittings that meet pipe grade steel requirements and for maintenance forms that document materials of a type-B full encirclement sleeve repair.

Note that a change in materials in an infrastructure requires a controlled process to manage the change in records if it is a large change in infrastructure (e.g., a re-route) or a small change (e.g., a single joint replacement). In either case, Data Miners will identify the locations in the existing pipeline information database that needs replacement. These segment(s) of pipe and fitting(s) will need to be retired into a controlled location according to the database schema (e.g., the protocol required by PODS, APDM, UPDM).

New "direct links" and new reliability rankings will be determined for the new records by the Data Miners according to the processes from Sections 4 and Sections 5. Ideally, the "direct links" to the old source records would persist with the retired segment(s) of pipe and fitting(s) to the retired location within the database (but this may not be possible with current database schema). The "direct links" to the old source record may not be allowed to remain with any new attribute data or system materials, because this will create a version control issue and the potential for misinterpretation by the company Decision Makers is too great.

As discussed earlier in the recommendations, normalization of records is an important step for enhanced data and recordkeeping for integrity management. On-going data collection represents the stage at which the industry can potentially create new projects with new records that repeat past mistakes by not meeting minimum quality and reliability criteria. Normalizing the records that should be produced for each reporting method can eliminate these mistakes but will require understanding and collaboration from those who create each record type and the Data Miners who review and assimilate them into the database.

6.3 Continuous Cross-Check and Quality Control

Operators can use measurements and observations from on-going field activities to gain confidence and fill gaps in the pipeline information database. This is accomplished by cross-checking new observations and measurements against the existing data within the database. This will require Data Miners to systematically review the new information as it comes in and take action based on its consistency to the database. These may be the same Data Miners who were conducting the document research in earlier sections.

The operator must track and document verifying measurements within a pipe population. A statistical reliability model and approach is presented in Appendix C to determine and track statistical confidence by verifying a data set. This model can be used as a metric to compare and prioritize confidence levels in different pipe populations based on the results of verifying measurements or observations that have been completed.

This metric could be used as part of an operator's plan, along with overall risk and the reliability of the source records, to prioritize verification activities (e.g., future field measurements). Pipe populations of lower confidence may be prioritized higher for verification efforts depending on the overall pipeline risk and the quality and reliability of the original source records from the pipeline information database at the specific location.

When comparing verification measurements to the pipeline information database, occasionally an operator will identify an inconsistency between the new information and the existing data within the database. The next section will describe managing and resolving these inconsistencies.

6.4 Isolating and Resolving Inconsistencies

It's inevitable when cross-checking that operators will obtain new information (e.g., through verification measurements, or surveys) that is inconsistent with a record in the pipeline information database. Making this type of discovery is the intent of incorporating cross-checks.

When an inconsistency is identified, it can present a challenge for operators, because it may not be possible to find a resolution without large-scale verification activities (e.g., running In-line inspection or performing many direct examinations). The following process is an effective way to systematically revisit the source records to see if large-scale verification activities are avoidable.

- Identify and isolate the data point(s) or source documents that are the cause of the records failure.
- Establish a boundary around the potential extents of the records failure.
- Systematically research and/or investigate within the established boundary and verify the data discrepancy until it is resolved.

These steps will allow an operator to identify the set of documents that may have caused the inconsistency issue. A detailed review of the documents related to the inconsistency will allow the operator to attempt to obtain a full understanding about the nature of the inconsistency.

The project numbers and geographic locations of the inconsistent data point are commonly used to identify the source records that are the cause of a records failure. During the review, the operator will search for trends that identify the initial set of source records and determine how the set should be pared down. There may be evidence that the records failure is specific to a certain type of pipe, a certain location, or project. The operator can establish a boundary (and reduce the boundary) around the extents of the records failure by reviewing the records for evidence that can reduce the set of records that require consideration. This is an important step because it creates efficiencies in the final step, which may require costly verification.

The final step is reviewing the applicable set of source records and attempting to find any new information possible to find what went wrong. The new information that led to the inconsistency may suggest more field work needs to be done, or it may lead to more specific document research that wasn't performed previously (e.g., the realization that a project file had been hidden away in an engineer's office). It is important to note that depending on each specific scenario, inconsistency, and set of source records, it may or may not be possible to obtain a full understanding from the source records alone.

If document review can't resolve the inconsistency, or if it becomes impractical to continue, the data that were the cause of the records failure should be tracked with the statistical reliability metric (See Appendix C). At the operator's discretion, the source records may need to be removed from the pipeline system information database (for appropriate locations and data points), which will create a gap. Either case will result in a downgrade in the system's records reliability, which will likely require verification by materials testing to return the system to its previous reliability levels.

This resolution can potentially be costly (depending on the final extent of the records failure). The industry is developing rapidly in materials testing technology. There are solutions currently evolving within in-line inspection, in-the-ditch methods that use proprietary tools and technologies, and in-the-ditch methods that can be applied more universally. An operator should consider a technology review to evaluate the most appropriate verification testing solution.

An example of an inconsistency that might be identified during cross-checking would be a Data Miner who, while reviewing an ultrasonic wall-thickness ILI report, observed a reported wall thickness of 0.308 inch for one joint in an area where the pipeline information database shows a nominal wall thickness of 0.365 inch. The Data Miner determined the ILI reported value of 0.308 inch was too low to be consistent with a nominal wall thickness of 0.365 inch.

The Data Miner used the following steps to help resolve this inconsistency.

- Identify and isolate the data point(s) or source documents that are the cause of the records failure.
- Establish a boundary around the potential extents of the records failure.
- Systematically research and/or investigate within the established boundary and verify the data discrepancy until it is resolved.

The Data Miner identified and isolated the source records that were the cause of the records failure by using the “direct link” to retrieve them from the GIS system (i.e., the source records linked to wall thickness at the location of the discrepancy). The source record was the as-built drawing from the original construction of the pipeline and it did not show the thinner-walled joint of pipe. The Data Miner established a boundary around the potential extents of the records failure by reviewing the remainder of the ILI report to ensure the wall thickness discrepancy did not occur in other locations.

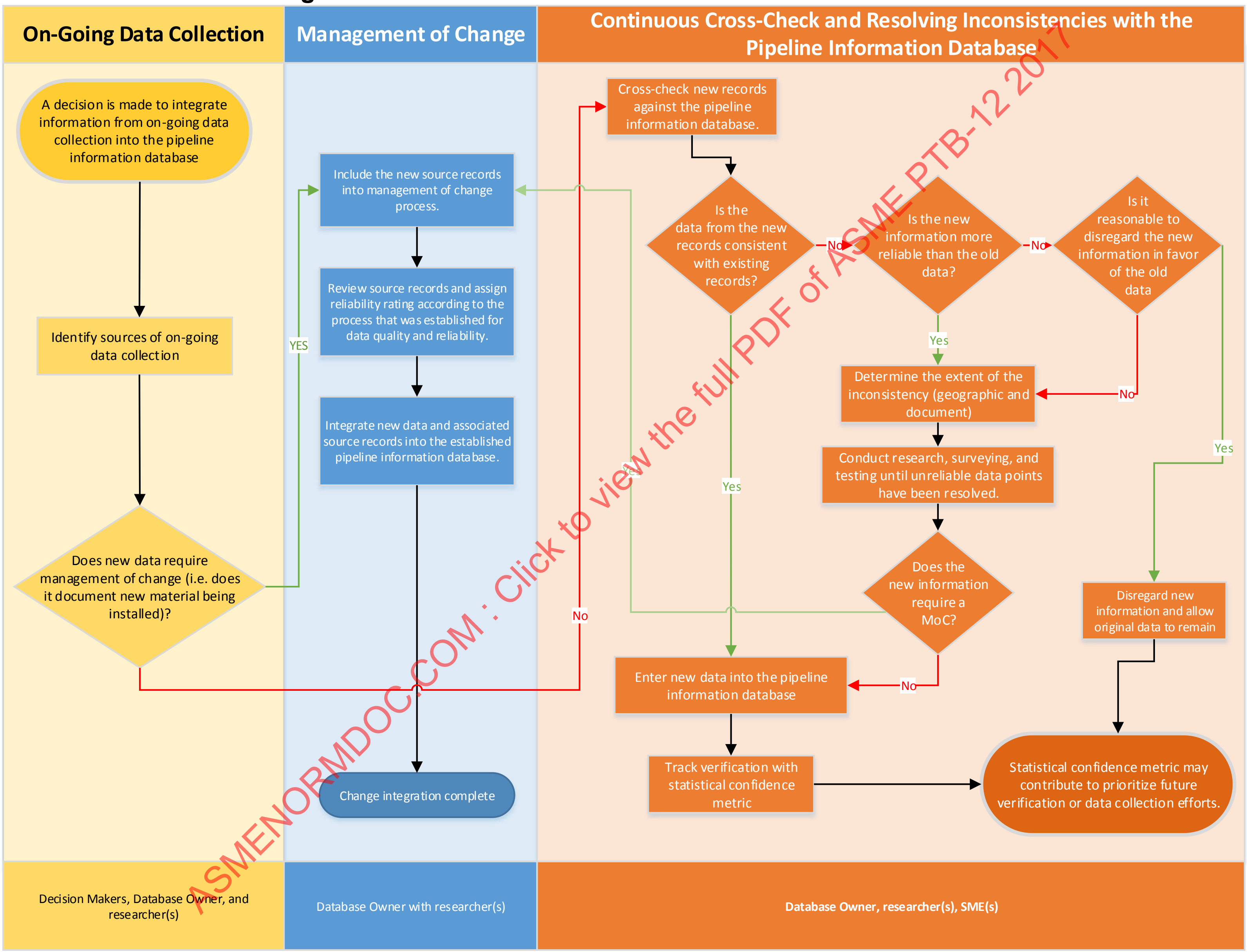
The Data Miner observed other locations with similar wall thickness, but they were all replacements performed following the previous in-line inspection. The Data Miner suspected this location was likely a replacement as well. The Data Miner continued to research but could not find the primary maintenance form documenting the repair. After reviewing the project files from the company network drive, he identified the project inspector and contacted him.

The project inspector remembered the exact excavation site and confirmed the replacement took place. The Data Miner received an email summary from the inspector and collected the project planning files from the network drive (i.e., material acquisition forms, one call, project log). The Data Miner provided his findings to the Database Owner, who wasn’t satisfied with the level of reliability that had been achieved, so an excavation was scheduled for verification purposes.

In-the-ditch measurements were taken for wall-thickness, hardness (which was used to estimate lower bound yield strength), and pipe chemistry. The direct examination report included these measurements along with observations about seam-type and coating. The Database Owner enlisted the help of a company SME in materials to compare the direct exam findings with the mill test results from the prior replacement project. The SME concluded the pipe chemistry and yield strength showed the pipe were likely the same type.

The Database Owner was satisfied with the level of reliability achieved with this new information. He instructed the Data Miner to initiate the management of change process for this location to replace the legacy source records with the new source records obtained and the inconsistency was resolved.

6.5 Work Flow Process Diagram



7 CLOSURE

In summary, the processes provided in Sections 3 through 6 of this guideline show an enhanced data and recordkeeping plan to manage records for pipeline integrity management.

Lessons learned from other industries reviewed in Section 1.2 have pointed to the following improvements that can potentially be made in the near term and longer term, including:

1. Normalize the data and record types that should be maintained.
2. Formally dedicate resources specifically to recordkeeping roles and responsibilities.
3. Provide more accessibility of the data and records to company deputies.

- The plan addresses the first lesson learned with a system to rank legacy pipeline system materials records with the document hierarchy in Section 5. The plan provides some normalization guidance for construction records for new pipe in Section 6.2. Normalization for the entire set of integrity management data is recommended for future work.
- The plan addresses the second lesson learned by indicating roles and responsibilities for each process, including Section 6 On-going Data Collection, which will require an on-going records maintenance team.
- The plan addresses the third lesson learned by defining a pipeline information database structure with direct links between the GIS interface and the verifying source records. This interface allows decision makers access to source records when making important fitness-for-service and risk decisions.

The guideline presented a detailed "plan for collecting, reviewing, and analyzing" and maintaining data that could be referenced in Sections 4.3 and 4.4 of ASME B31.8S. The plan implements a philosophy that was created from the records review initiative following San Bruno to create a structured database and make source records (as the basis for the information in the database) accessible to company Decision Makers.

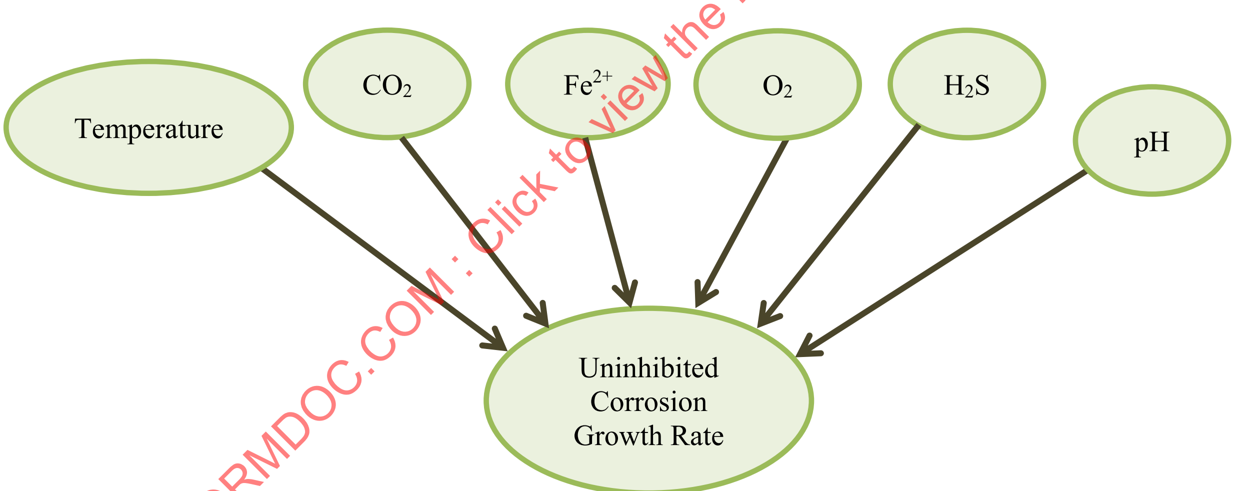
APPENDIX A: GUIDANCE FOR USING A BAYESIAN NETWORK TO BOUND UNKNOWN VALUES INTO LIKELY VALUES

The following example of a Bayesian network was used, with the authors' permission, from "Quantitative Assessment of Corrosion Probability – A Bayesian Network Approach," NACE International, Corrosion Vol. 70, No. 11, 2014. [35]

Bayesian networks have proven useful to pipeline operators for filling unacceptable gaps in pipeline datasets (or databases). One of the Bayesian methods to make this determination could be to build a Bayesian Network in order to calculate the probability of different states. This method is most effective if there are multiple known variables that have an implied influence on the unknown (gap) variable. This type of analysis has been used for applications in the pipeline industry during risk determination, root cause analysis, or maintenance prediction.

A Bayesian Network is a type of graphic statistical model that represents probabilistic relationships between variables. This analytic process is used to model the probabilities of cause and effect relationships to make predictions based on known distributions and values. If parameter "X" is true, then the network shows the probability of the values for related parameters "W," "Y," and "Z." Below is an example of the graphical model for a simple Bayesian network.

Figure A-1: Example Graphical Bayesian Network Model for Uninhibited Internal Corrosion Growth Rate



In this example, the goal is to determine the uninhibited internal corrosion growth rate (i.e., the unknown "gap" variable) for an area taking into account environmental information.

1.1 Creating a Bayesian Network

A functional Bayesian Network can be created more quickly and easily with a software package specifically designed to streamline the process. There are various commercial and free options with different levels of complexity and functions. The commercial options will be the most comprehensive but might not be practical if there are no plans for repeat use. Software options can be found from various websites, and it is recommended to look at the available resources online and select the best option for the desired application. If there are no suitable free options, it is suggested to consult with an industry expert with commercial Bayesian Network software experience that preferably also has pipeline experience.

1.1.1 Developing the Graphical Model

The following instructions are representative of typical software options, but if the instructions below are not helpful, then it is recommended to seek out instructional material for the specific software selected. Two components are needed to create a basic Bayesian Network: the “Chance” bubble depicted as an oval and the “Arc” connectors depicted as an arrow. The “Chance” ovals are used to represent variables and the “Arc” arrows are used to assign the relationships between the different “Chance” ovals.

Begin by placing “Chance” ovals and assigning variable names to each one. This step requires a detailed understanding of the result and the factors that are directly influencing it. The model is more likely to be an accurate representation of reality if most (or all) of the influencing variables are present and are considered. It is important the arrows that link the variables to one another in the model reflect direct cause and effect relationships and not simply a correlation. For example, when looking at pipeline failure data, there may be a correlation between the year the pipe was manufactured and the number of failures from that decade. The age of the pipe does not directly cause the pipe to fail, but it does directly influence some of the pipe properties, such as the coating and seam type. To correctly link the manufacturing year to the failure in a Bayesian Network, there are intermediate variables needed, such as coating and seam type.

1.1.2 Developing the Conditional Probability Tables

Once all variables and relationships are created and assigned, the user must input the probabilities for each case the variable could represent. The case could represent a decision (e.g., yes or no), a numerical value or range, or a current state (e.g., on or off). The relationships in the Bayesian network become more complex with each additional case that the variables could potentially represent.

The probability for each case can be assigned manually or, for networks containing variables with multiple cases, using a program to calculate the probabilities for each case. After all the cases are assigned probabilities, all values in one column (i.e., possible cases for a single variable) must add up to one. In most software there is “normalize” option when entering data to scale all the values proportionally so that they automatically add to one. Although this will create a probability set that is functional, this does not fix incorrect data entries. A Bayesian Network will only be as accurate as the probabilities assigned to different variable cases and the relationships assigned, so it is pivotal that these be verified once completed.

There are three main ways to determine appropriate probabilities and relationships: SME input, research/observation data, tested and recognized models. For variables with many potential cases, SME input may not be the best information source. However, SME input is highly useful when assigning relationships and for identifying scenarios where the condition probability will be either zero or one or identifying scenarios when using a uniform distribution is appropriate.

Research or observed data can be applied to determine conditional probabilities. The quality of this information is only as good as the source and sample it was taken from. If an operator wanted to determine the probability distribution for pipe grade, the operator’s total pipe mileage could be broken down into percentages based on the grade and then applied to the Bayesian Network. Depending on the variable, there may also be industry sources or studies that will provide the necessary data to determine percentages.

Using a model to determine conditional probabilities is typically the most preferred method, when applicable. It is also the ideal method when a variable has multiple relationships, which will result in a large table of conditional probabilities. This method determines probabilities from the results of running the model based on different inputs each with a given range, similar to a Monte Carlo calculation. By calculating possible combinations, a probability distribution can be developed by creating bins and determining the percentage of results that fell inside the bin. Some pipeline related models could include something as complex as multiphase flow or corrosion rate or something well established, such as the design pressure equation.

Take the corrosion growth example depicted above. Once the graphical model is created, the probabilities will need to be assigned. The following tables show potential probability assignments for this Bayesian Network assuming none of the input parameters are known.

Table A-1: Probability Assignments for Example Bayesian Network

T		CO_2		Fe^{2+}		O_2		H_2S		pH	
C	%	Bar	%	ppm	%	ppb	%	ppm	%	—	%
20-40	25	0-0.1	25	0-10	33	0-10	25	0-10	25	4-5	25
40-60	25	0.1-1	25	10-50	33	10-100	25	10-100	25	5-6	25
60-80	25	1-10	25	50-100	33	100-1,000	25	100-1,000	25	6-7	25
80-100	25	10-100	25			1,000-10,000	25	1,000-10,000	25	7-8	25

For the top row of variables, each one was split into appropriate ranges as determined by SMEs and assigned uniform probability distributions. It is possible to go in later and update these probabilities based on known information, typically called "Evidence" by various software packages. There could also be additional levels of detail added to make these distributions different, such as temperature is affected by barometric pressure and the time of year and the pH of the soil is impacted by the location. Adding these extra levels of detail will change these variables from uniform distributions to a custom distribution that will need to be determined by a model or quality data and potentially improve the accuracy of the model. However, it is possible that these extra variables will have minimal impact on the desired result or can be compensated for by known information, so it is best to consult with an SME to determine the level of detail required.

To determine the probability distribution for the corrosion growth rate variable, a corrosion rate model that considers these inputs will need to be used due to the number of potential combinations. There are six variables, five with four possible cases and one with three possible cases, that impact corrosion growth. This will result in 3072 different probabilities per corrosion growth bin because each combination of variables will need to be considered. Below is a table containing the first and last five columns of the conditional probability table for the corrosion growth variable. In this case, the corrosion rate probabilities are based on a corrosion model that uses the environmental parameters as input.

Table A-2: Example Results from the Conditional Probability Table

Temperature	Celcius	20-40	20-40	20-40	20-40	20-40	…	80-100	80-100	80-100	80-100	80-100
CO_2	Bar	0-0.1	0-0.1	0-0.1	0-0.1	0-0.1	…	10-100	10-100	10-100	10-100	10-100
Fe^{2+}	ppm	0-10	0-10	0-10	0-10	0-10	…	50-100	50-100	50-100	50-100	50-100
pH	—	4-5	4-5	4-5	4-5	4-5	…	7-8	7-8	7-8	7-8	7-8
O_2	ppb	0-10	0-10	0-10	0-10	10-100	…	100-1,000	1,000-10,000	1,000-10,000	1,000-10,000	1,000-10,000
H_2S	ppm	0-10	10-100	100-1,000	1,000-10,000	0-10	…	1,000-10,000	0-10	10-100	100-1,000	1,000-10,000
Corrosion rate probability in mm/year	0-0.01	0	0	0	0	0	…	0	0.65	0	0	0
	0.01-0.1	0.065	0.525	0.42	0	0.08	…	0	0.35	0	0	0
	0.1-1	0.865	0.475	0.58	1	0.86	…	0	0	0	0	0
	1-2	0.07	0	0	0	0.06	…	0.16	0	0.8	0.57	0.185
	2-5	0	0	0	0	0	…	0.84	0	0.2	0.43	0.815
	5-10	0	0	0	0	0	…	0	0	0	0	0
	10-20	0	0	0	0	0	…	0	0	0	0	0

1.2 Using a Bayesian Network

Once the network has been created and all the inputs have been verified, it can be used within the Bayesian Network software package, by referencing the model using an external program routine, such as Excel, Visual Basic (VBA), or MATLAB, or by using a pre-existing software.

1.2.1 Running in the Bayesian Network Software

To run the network in the software that was used to create it, first flip the "Chance" ovals to view as a Bar Chart. Some packages may have this as a default, but others will have it as a secondary view which can typically be found under a "View" or "Node Format" related menu. Some packages may use other methods or not have this functionality, so it is recommended to review instructional materials for the specific software used if the suggestions above are not helpful.

Once the different variables are shown as a bar chart, make sure to update the view as some packages do not update automatically. This can usually be found as an icon along the top tool bar. After the page is refreshed, the bar charts should display the probabilities for each case of each variable.

After updating the network in the software, the table below shows the results with no known data.

Table A-3: Bayesian Example Results for Scenario with No Known Data

T		CO_2		Fe^{2+}		O_2		H_2S		pH		Corr. Rate	
C	%	Bar	%	ppm	%	ppb	%	ppm	%	—	%	mm/year	%
20-40	25	0-0.1	25	0-10	33	0-10	25	0-10	25	4-5	25	0-0.01	3
40-60	25	0.1-1	25	10-50	33	10-100	25	10-100	25	5-6	25	0.01-0.1	15
60-80	25	1-10	25	50-100	33	100-1,000	25	100-1,000	25	6-7	25	0.1-1	43
80-100	25	10-100	25			1,000-10,000	25	1,000-10,000	25	7-8	25	1-2	18
												2-5	8
												5-10	4
												>10	9

This step is also a good chance to verify that the conditional probabilities yield accurate results. Sometimes the initial distributions may not be intuitive to conceptualize (which is the advantage of using a Bayesian Network), so it is important to evaluate the network using some known data points. To enter known information, double click the case in the bar chart view and the probability will change to 100%. It is also possible to set "uncertain" evidence, where the updated probabilities eliminate one or more of the states for a variable. Doing this does not erase the data previously entered and can be undone at any time. It is also important to update the view if the network does not auto update. In the example tables below, the red cells are known values, the blue cells are uncertain values, and the green cells are unknowns and are recalculated by the Bayesian Network.

Table A-4: Bayesian Example Results for Scenario with a Mixture of Known, Unknown, and Uncertain Data

T		CO_2		Fe^{2+}		O_2		H_2S		pH		Corr. Rate	
C	%	Bar	%	ppm	%	ppb	%	ppm	%	—	%	mm/year	%
20-40	100	0-0.1	0	0-10	33	0-10	50	0-10	50	4-5	0	0-0.01	0
40-60	0	0.1-1	0	10-50	33	10-100	50	10-100	50	5-6	100	0.01-0.1	0
60-80	0	1-10	100	50-100	33	100-1,000	0	100-1,000	0	6-7	0	0.1-1	45
80-100	0	10-100	0			1,000-10,000	0	1,000-10,000	0	7-8	0	1-2	21
												2-5	27
												5-10	7
												>10	0

Table A-5: Bayesian Example Results for Scenario A with a Mixture of Known and Unknown Data

T		CO_2		Fe^{2+}		O_2		H_2S		pH		Corr. Rate	
C	%	Bar	%	ppm	%	ppb	%	ppm	%	—	%	mm/year	%
20-40	25	0-0.1	0	0-10	33	0-10	0	0-10	0	4-5	0	0-0.01	0
40-60	25	0.1-1	100	10-50	33	10-100	100	10-100	0	5-6	100	0.01-0.1	0
60-80	25	1-10	0	50-100	33	100-1,000	0	100-1,000	0	6-7	0	0.1-1	100
80-100	25	10-100	0			1,000-10,000	0	1,000-10,000	100	7-8	0	1-2	0.6
												2-5	0
												5-10	0
												>10	0

As more precise information is added to the network, unknowns will become more refined. Also, as demonstrated in the second table, although some variables remain as unknowns, a corrosion growth rate might still be able to be determined with a high degree of certainty.

In the example, the corrosion growth rate is most likely the desired information, but it is important to still think of a Bayesian Network as a loop rather than as a routine with a start and end. For example, if the corrosion growth rate was already known, the Bayesian Network can calculate the probabilities for the other unknown variables. To do this in the creation software, double click the bin case on the bar chart representation of the corrosion growth variable and update the view if necessary. Below is an example using a known corrosion growth rate, the cell color meanings are the same as the above examples.

Table A-6: Bayesian Example Results for Scenario B with a Mixture of Known and Unknown Data

T		CO_2		Fe^{2+}		O_2		H_2S		pH		Corr. Rate	
C	%	Bar	%	ppm	%	ppb	%	ppm	%	—	%	mm/year	%
20-40	0	0-0.1	0	0-10	0	0-10	0	0-10	0	4-5	100	0-0.01	0
40-60	0	0.1-1	0	10-50	0	10-100	50	10-100	0	5-6	0	0.01-0.1	0
60-80	100	1-10	100	50-100	100	100-1,000	50	100-1,000	0	6-7	0	0.1-1	0
80-100	0	10-100	0			1,000-10,000	0	1,000-10,000	100	7-8	0	1-2	0
												2-5	100
												5-10	0
												>10	0

This functionality of entering in some known/uncertain information to determine the unknown values is the intended goal of using a Bayesian Network to make decisions for the “most likely case” and “worst probable case” values in the bounding likely values process.

1.2.2 Running in Other Software

The authors of this guideline don’t necessarily expect software will be required for all Bayesian Network applications resulting from the practices in this guideline. However, once a Bayesian Network is built it can also be used by custom programs/macros written in programs such as MATLAB or Excel. An API or library of commands compatible with Bayesian Networks can be found for free or from a paid provider online (e.g., SMILE from BayesFusion, LLC [32]) and will facilitate use of the Bayesian Network. Also available are premade software to handle and visualize the Bayesian Network results.

1.3 Hypothetical Case Study

During an internal audit process an operator identified missing wall thickness and SMYS values along one of its pipeline segments. At some locations, just one variable was unknown, and at other locations both were unknown. Without accurate information, the pipeline is at risk for inaccurate fitness-for-service evaluations and incomplete design data for MAOP calculations. With these considerations, the operator assembled a small team to compile and research relevant records to compile materials data of the other representative and potentially comparable pipe types within the system.

The team was able to compile records about the materials information for other lines constructed during the same time period. There were also some design pressure calculations found that confirm the operating pressure. The team realized that due to multiple unknown values, it needed to use a process that could handle a large amount of combinations that can handle multiple factors or variables. The team decided that the best course of action forward is to build a Bayesian Network to determine the missing wall thickness and SMYS values.

First, the Bayesian Network team identified that the missing variables were most commonly associated through calculating the design pressure. The team referred to the internal design pressure formula listed in CFR 195.106[9] to start compiling a list of nodes.

$$P = \left(\frac{2St}{D}\right) * E * F$$

Where,

P = Internal design pressure

S = Yield strength

t = Nominal wall thickness

D = Nominal outside diameter

E = Seam joint factor

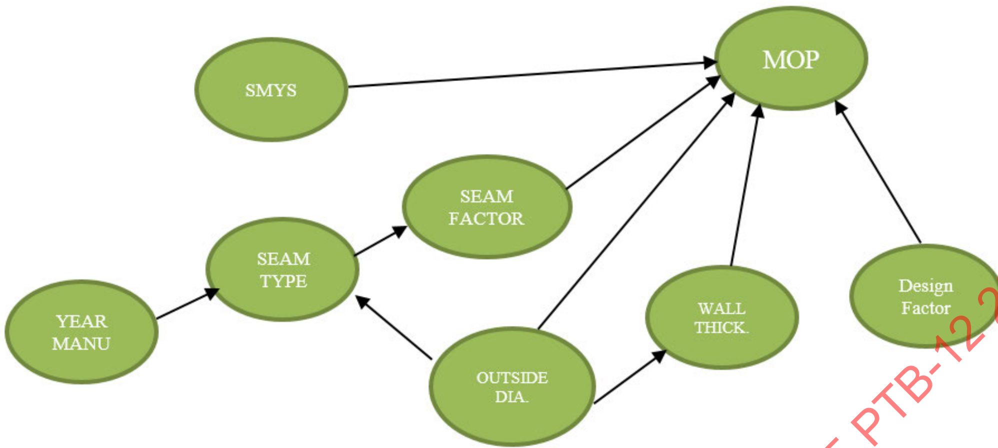
F = Design factor

The team began to then list all possible values for SMYS, wall thickness, diameter, seam factor, and design factor. During this process, the team also discovered that certain properties were also impacted by the year manufactured and the seam type so they were added as extra nodes to network. The team identified 11 possible SMYS values, 51 possible wall thickness values, 19 possible outside diameters, three possible seam factors, three possible design factors, ten possible decades, and eight possible seam types. Then to list out possible design pressures, the team elected to use pressure bins for every 10 psig starting at 0 psig and going to 2000 psig.

After establishing the nodes and entering in all the possible values, the team began assigning the relationships. The figure below shows the resulting network diagram.

[9] This equation is similar to other regulatory design formulae, such as CFR 192.105 and Z662 4.3.3.1.1.

Figure A-2: Example of a Network Diagram that shows the relationships between nodes

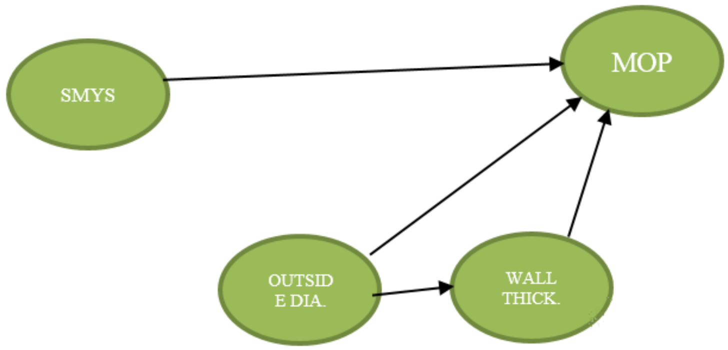


These relationships were determined based on the design pressure model as well as some of the research documents used. Combinations of SMYS, seam, outside diameter, and wall thicknesses were identified from manufacturing standard API 5L. After studying the API 5L combinations in conjunction with the legacy procurement practices, pipe grade A and a number of seam weld types were eliminated from possibility. The only relationship remaining from this source was between the outside diameter and wall thickness. Also, CFR 195.106 specified a simple relationship between the seam type and seam factor. Using these extra relationships will allow the team to refine the network.

The next step was for the team to start creating the conditional probability tables. Based on this network, tables were needed for Seam Type, Seam Factor, Outside Diameter, Wall Thickness, and MOP nodes. Most of these tables were simple to create as there was only one relationship; and most of these relationships were in a "yes"/"no" format (e.g., "yes," 4" OD pipe was made with a 0.250" WT, but "no" it was not made with 1" WT). This resulted in a small number of table cells to fill in, a uniform distribution for all "yes" pairs, and zero for all "no" pairs.

Next, the team addressed the conditional probability table for the "MOP" node. This table was the most complex and most time-consuming part of the network development because this cell had five different relationships (one for each path leading into the "MOP" node). Each combination of possibilities was considered to determine the distribution of possible design pressures. In total the team needed to consider 95,931 possible combinations of the five variables in the design pressure model to complete the Bayesian network. Completing this Bayesian network was possible with the right software, but by using the observations made during research of API 5L the team was able to refine the network instead. Two of the nodes (seam type and design factor) were replaced with constant values, which effectively removed them from the model and reduced the size and complexity significantly. The total number of property combinations was reduced to 10,659 and the amount of time and resources to set-up the software and write the problem-solving script was also reduced. The figure below shows the simplified network.

Figure A-3: Example of a Simplified Network Diagram that shows the relationships between nodes



When preparing the problem-solving script (e.g., in VBA) it was important to ensure that the order of the pipe properties used to calculate the "MOP" in the script appeared in the same as the order as they appeared in the Bayesian Network software. The tables could then be copied directly from Excel into the Bayesian Network software.

Once all the probability tables were created, the model was ready. The line was operated at a pressure of 985 psi and had a nominal outside diameter of 12". These values had been verified by multiple sources collected by the research team and were consistent with the current operating parameters. The team set these parameters as known within the Bayesian Network. After the evidence was set, the network revealed that grade B SMYS and 0.172" WT were the worst-case values for SMYS and wall thickness, respectively (but not as combination), and X42 SMYS and 0.250" WT was the most likely combination. The team used these values going forward in the missing value verification process and continued with the process detailed in Section 4.

It is important to note that in this case, the operator had a high level of confidence in its design practices for the pipe in question. There can be inherent flaws using the current operating pressure to make inferences about material properties and this step may not be recommended in all circumstances. Each operator must decide when this method (as with other Bayesian input variables) is appropriate for purpose.

APPENDIX B: GUIDANCE FOR USING A DOCUMENT HIERARCHY TO SYSTEMATICALLY DETERMINE DOCUMENT RELIABILITY

The discussion in Section 7 focused on establishing a document hierarchy, or document ranking scheme, which can be used to determine the quality and reliability of the data extracted from within each source record. Once created, the document hierarchy not only establishes data quality and reliability, it also provides an effective tool to assist in efficient data mining. This can become essential if a large data mining effort is required.

A team consisting of Data Miners with oversight from engineers is typically assembled to complete large data mining efforts, like those that occur when construction records are reviewed to verify maximum allowable operating pressure (MAOP). This team structure can provide cost advantages because the data mining team members are typically more cost-effective (less expensive) than engineers.

The Data Miners in this team structure require oversight from engineers and process controls to assure the quality of their results. The document hierarchy is a valuable process control for this type of effort. It can create efficiencies by helping the Data Miners to understand which document types are more likely to contain the targeted information. It can also create repeatability by providing guidance about interpreting document precedence and how to resolve conflicting information.

This appendix discusses how to use a document hierarchy during a data mining project, and it provides a brief hypothetical case study to convey a potential real life context.

Data Mining with a Document Hierarchy

When functioning properly, a document hierarchy will allow Data Miners to target higher value documents and help them to discern information between conflicting source records to arrive at the most reliable set of pipeline data. The authors have typically used document hierarchies when data mining construction records, pressure test, and maintenance records, but have also found they apply to in-line inspection results and expect that they apply to other data types as well (e.g., potentially control room pressure records) once properly configured and found fit for purpose by company SMEs who are familiar with the targeted data and the sources under review.

The figure below shows a fully developed "Method 1" document hierarchy based on document type. This document hierarchy is fit for purpose when the source records under review are very consistent. This has been observed to occur when common procedures were followed closely and common forms were used consistently over long periods of time. This practice produces document types that are very consistent between projects and can be relied upon to have the same types of data.

A Data Miner would use the document hierarchy in conjunction with a digital document database and project list to target and retrieve high value documents on a project by project basis to reconstruct each pipeline. Here are a few steps the Data Miner might follow:

1. Once the project numbers are organized chronologically for a particular pipeline, the Data Miner will use the document index metadata to target "mapping documents" (like drawings) from the oldest project to define the project extents (with GIS stationing, if available).
2. Once the Data Miner understands where the project took place he will review the available highest ranking mapping documents to identify where various pipe and fitting materials were installed along the line.
3. If the highest-ranking mapping document provides adequate attribute information for the Data Miner's research requirements, the Data Miner moves to the next pipe, fitting, etc., but if the mapping document has gaps in attribute data or requires a second verifying document, the Data

Miner continues to review documents from the same project number. The Data Miner will use the metadata to target the highest value documents first. If there is a "Design Materials Engineering Document Report" (from Table B-1, below) available in this project file, that will be targeted first to fill the materials data gaps. This search will continue until all gaps are filled or project records are exhausted.

4. The Data Miner will then move to the next oldest project on the project list and repeat this process. If projects are found to replace one another, the materials information will be treated in a similar manner to the MOC process from Section 6.2. When all projects have been reviewed, the pipeline is current and the on-going data collection begins.

Table B-1: Fully Developed Method 1 Document Hierarchy Based on Document Type

Ranking	Source Record Type	Number of Required Documents
1	Design Materials Engineering Report	1 required
2	Completion Drawing	1 required
3	Material Certifications	1 required
4	Pressure Test Job Report	1 required
5	Welding Inspection Report	1 required
6	Materials Invoice	2 required
7	Material Purchase Order	2 required
8	Material Transfer Order	2 required
9	Re-Claimed Material	2 required
10	Atlas Sheets	2 required
11	Strip Maps	2 required
12	Material Requisition	2 required
13	Construction Bid Specifications	2 required
14	Construction Inspector Notes	2 required
15	Field Notes	2 required
16	Notes from Field	2 required

Resolving Conflicting Information with a Document Hierarchy

Data Miners will also use the document hierarchy to resolve conflicting information. The steps below provide an example of how a Data Miner might use a document hierarchy to resolve conflicting information.

1. A Data Miner has defined project extents for the current project under review and researched pipe and components drawn on a strip map that was used to document the materials installed. The strip map shows grade X-42 pipe with 0.365-inch wall thickness, but the document hierarchy requires a second verifying document for a strip map so the Data Miner continues researching.
2. The Data Miner opens a "construction bid specification" that shows the pipe with 0.365-inch wall thickness with grade B (which is in conflict with the "X-42, " mentioned above). In order to resolve this conflict the Data Miner must continue to research.

3. The Data Miner reviews a Material Invoice that shows the 0.365-inch wall thickness pipe for the project was X-42, in agreement with the strip map. The Data Miner consults the document hierarchy, which specifies the purchase order and strip map both rank higher than the construction bid specification.
4. The Data Miner uses the document ranking order to override the construction bid specification and verify the grade X42 with the strip map and material invoice.

The examples above were both simplified to show the principle of the document hierarchy. Method 1 is applicable for document sets that have a high degree of consistency, but they will still need to be reviewed for reliability flags (e.g., missing signatures, date/quantity inconsistencies, etc.) and the review may not be as simple as described. In practice, Data Miners will need to be trained to scrutinize the reliability of documents in addition to using the document hierarchy as a guidance tool, which is why engineer oversight and review is critical.

The examples described above could also apply to a Method 2 document hierarchy, but the Data Miner would have to perform more scrutiny of each document and interpretation of the documents' attributes (since a Method 2 document hierarchy is typically used for more inconsistent document sets).

When implementing a document hierarchy into a data mining process for the first time, a low production pilot is highly recommended. Unexpected document types and procedural gaps are common at the beginning of data mining projects. Taking the time to conduct a low production pilot (with 1-3 Data Miners for 3-6 weeks) will allow a team to address these challenges prior to wasting resources during a full-production effort.

APPENDIX C: STATISTICAL METHOD FOR TRACKING PIPELINE DATA CONFIDENCE AND RELIABILITY

This appendix outlines an approach to calculate the number of digs needed to verify a material property such as pipe grade, or wall thickness, is consistent with verification measurements to a given statistical confidencel level on a given pipeline. This calculation can also be used as a reliability metric when comparing verifiability for prioritization between different pipe populations.

The approach is based on establishing a lower bound likelihood that any member of a population is consistent with results obtained by (prior) sampling of the population. A sample is consistent if it matches prior results, and it is inconsistent if it does not. The approach can be used, for example, to estimate a lower-bound likelihood any random pipe joint in a pipeline is the same grade or wall thickness as that found by sampling one or more locations.

In this appendix, a population is a finite set of something (e.g., pipe joints of unknown type or wall thickness on a pipeline). Sampling is the process of selecting a random member of the population to determine if it meets predefined criteria (e.g., if the pipe joint is X52). Consistency is the proportion of the population that matches the predefined criteria.

The approach is based on a hybrid algorithm that combines an exact solution for an infinite population with a newly developed finite-population normal approximation based on the Agresti-Coull method in API 1163. [36] Any normal approximation requires a reasonable sample size to be accurate. The exact method (detailed in this appendix) is conservative for finite populations. By using a hybrid methodology, the finite population approach will be used once sufficient number of joints have been assessed. For smaller numbers of joints the exact conservative solution may be used.

Binomial proportions take on one of two states (e.g., sampled yield strengths are or are not consistent with a given grade). The confidence interval surrounds the proportion of pipe that matches a target property in a finite population of pipe joints. The method can be used in situations where the sampling results are not consistent.

As an example of the levels of confidence that are achievable by completing a very large number of verification measurements, Table C-1 summarizes the number of digs needed to achieve a given confidence level and consistency for an infinite population using the exact solution, assuming no inconsistencies. Here, consistency is defined as the minimum (lower bound) proportion of pipe on the pipeline that matches a target grade. If an operator wants to have 95% confidence that 95% of the pipe joints on a pipeline are the same grade, he must perform 59 digs, all of which must be consistent with the target grade. If sufficient data is available, the finite population approach covered later may require fewer samples.

Table C-1: Number of Digs Required as a Function of Confidence Level and Consistency with the Exact Conservative Solution

		Consistency			
		80%	90%	95%	99%
Confidence Level	80%	8	16	32	160
	90%	11	22	45	228
	95%	14	29	59	297
	99%	21	44	90	456

In the event inconsistencies cannot be explained by typical variability in various pipe characteristics (e.g., one pipe joint is X52 and another is X65), separate populations must be developed. This scenario typically coincides with multiple inconsistent measurements.

Statistical Models for Calculating Confidence Levels

This section discusses a procedure to calculate a lower bound "consistency" (i.e., proportion of pipe meeting a given criteria). The lower consistency bound is for the full population (relevant pipe segment with assumed similar diameter, wall thickness, grade, seam, install date, manufacturer, and manufacture date) based on the available sample information.

<u>Exact Binomial Confidence Interval Formulas</u>

This section is primarily taken from Harper (2005) [37], including the Excel VBA function that was used for the exact results shown in this appendix. Hollander and Wolfe [38] explain exact Clopper-Pearson [39] binomial confidence limits. The lower and upper confidence bounds are given below.

$$p_L^{\alpha}(n,B)=\frac{B}{B+(n-B+1)\,f_{\alpha/2,2(n-B+1),2B}}$$

$$p_U^{\alpha}(n,B)=1-p_L^{\alpha}(n,n-B)$$

where B is the number of successes in the n Bernoulli trials and f_{γ,n_1,n_2} is the upper γ^{th} percentile of the F distribution with n_1 and n_2 degrees of freedom.

Table B-2, which is based on a 95% confidence level, provides an example to aid in the interpretation of lower consistency bounds. Table B-2 shows that the full population has <u>at least</u> (similar to a worst case) a given consistency level. Consistency may be viewed as either a proportion in the range [0, 1] or a percentage in the range [0%, 100%]. The computed consistency is a function of:

- Desired confidence level (80%, 90% or 95% are suggested)
- # observations (labeled "# joints inspected"): these are the row labels in the first column
- # inconsistencies: these are the column labels in row 10.

Table B-2: Lower-Bound Consistency Table at a 95% Confidence Level for an Infinite Population

# Joints inspected	# Inconsistencies 0	1	2	3	4	5	6	7	8	9
1	0.050	0.000								
2	0.224	0.025	0.000							
3	0.368	0.135	0.017	0.000						
4	0.473	0.249	0.098	0.013	0.000					
5	0.549	0.343	0.189	0.076	0.010	0.000				
6	0.607	0.418	0.271	0.153	0.063	0.009	0.000			
7	0.652	0.479	0.341	0.225	0.129	0.053	0.007	0.000		
8	0.688	0.529	0.400	0.289	0.193	0.111	0.046	0.006	0.000	
9	0.717	0.571	0.450	0.345	0.251	0.169	0.098	0.041	0.006	0.000
10	0.741	0.606	0.493	0.393	0.304	0.222	0.150	0.087	0.037	0.005
11	0.762	0.636	0.530	0.436	0.350	0.271	0.200	0.135	0.079	0.033
12	0.779	0.661	0.562	0.473	0.391	0.315	0.245	0.181	0.123	0.072
13	0.794	0.684	0.590	0.505	0.427	0.355	0.287	0.224	0.166	0.113
14	0.807	0.703	0.615	0.534	0.460	0.390	0.325	0.264	0.206	0.153
15	0.819	0.721	0.637	0.560	0.489	0.423	0.360	0.300	0.244	0.191
16	0.829	0.736	0.656	0.583	0.516	0.452	0.391	0.333	0.279	0.227
17	0.838	0.750	0.674	0.604	0.539	0.478	0.420	0.364	0.311	0.260
18	0.847	0.762	0.690	0.623	0.561	0.502	0.446	0.392	0.341	0.291
19	0.854	0.774	0.704	0.641	0.581	0.524	0.470	0.418	0.368	0.320
20	0.861	0.784	0.717	0.656	0.599	0.544	0.492	0.442	0.394	0.347
21	0.867	0.793	0.729	0.671	0.616	0.563	0.513	0.464	0.417	0.372
22	0.873	0.802	0.741	0.684	0.631	0.580	0.532	0.485	0.439	0.395
23	0.878	0.810	0.751	0.696	0.645	0.596	0.549	0.504	0.460	0.417
24	0.883	0.817	0.760	0.708	0.658	0.611	0.565	0.521	0.479	0.437
25	0.887	0.824	0.769	0.718	0.670	0.625	0.580	0.538	0.496	0.456
26	0.891	0.830	0.777	0.728	0.682	0.637	0.595	0.553	0.513	0.474
27	0.895	0.836	0.785	0.737	0.692	0.649	0.608	0.568	0.529	0.491
28	0.899	0.841	0.792	0.746	0.702	0.661	0.620	0.581	0.543	0.506
29	0.902	0.847	0.798	0.754	0.712	0.671	0.632	0.594	0.557	0.521
30	0.905	0.851	0.805	0.761	0.720	0.681	0.643	0.606	0.570	0.535

For example, an operator conducts 14 excavations, all of which are consistent with a given grade. As seen in Table B-2, a lower bound consistency of 0.807 (80.7%) is obtained at the 95% confidence level when zero inconsistencies are found in 14 observations. A consistency of 0.807 implies that the worst case (or 95% lower bound) consistency for the full population is at least 80.7%. That is, at least 80.7% of the population is the same grade. Of course, for this example it could be as high as 100% since no inconsistencies have been found.

If the operator wants a higher statistical consistency, more excavations are required. For example, if the operator wants 90% of the population to be the same grade, 29 consistent excavations are needed. What happens if the operator finds an inconsistent measurement? The number of required digs to achieve the same statistical consistency increases. If there is one inconsistency, Table B-2 shows that 22 excavations are needed to conclude that 80% of the population is the same grade. As the number of inconsistencies increases, the consistency drops for a given number of excavations.

Table B-3, Table B-4, and Table B-5 show similar lower-bound consistency values at the 99%, 90%, and 80% confidence level.