SASIG Long Term Archiving & Retrieval of Digital Product Definition Data

Format Recommendation

AIAG PUBLICATIONS

An AIAG publication reflects a consensus of those substantially concerned with its scope and provisions. An AIAG publication is intended as a guide to aid the manufacturer, the consumer, and the general public. The existence of an AIAG publication does not in any respect preclude anyone from manufacturing, marketing, purchasing, or using products, processes, or procedures not conforming to the publication.

CAUTIONARY NOTICE

AIAG publications are subject to periodic review and users are cautioned to obtain the latest editions.

MAINTENANCE PROCEDURE

Recognizing that this AIAG publication may not cover all circumstances, AIAG has established a maintenance procedure. Please refer to the Maintenance Request Form at the back of this document to submit a request.

Published by:

Automotive Industry Action Group

26200 Lahser Road, Suite 200

Southfield, Michigan 48033

Phone: (248) 358-3570 • Fax: (248) 358-3253

APPROVAL STATUS

The AIAG Materials Management Steering Committee and designated stakeholders approved this document for publication on December 12, 2006.

ISBN#:

FOREWORD

Many companies are migrating or have migrated their product definition and lifecycle management authoring processes from traditional hard-copy, paper based document management processes to processes that highly leverage computer aided/digital information creation techniques. As a consequence of this activity, new processes must also be defined to archive digital information and preserve access to it, in compliance with business and regulatory requirements.

Certain classes of product definition data specify multi-decade retention periods. Over these time periods, changes in both the editing and storage technologies impact an organization's ability to retrieve and use product information. All organizations which use digital product information will need strategies and processes that maintain the usability of the information over multiple generations of technology.

The SASIG Long Term Archiving & Retrieval Project is developing a set of recommendations to guide companies to effective and efficient archival and retrieval practices. The recommendations are partitioned into four topic areas: 1) Format, 2) LTAR Process, 3) Retention Time Periods, and 4) Quality Assurance.

This document addresses the set of format recommendations. In particular this document aims to provide a company with the key functional and technical elements to consider when choosing an archival format(s). To aide in this decision making process, this recommendation includes a classification scheme for understanding of the different types of formats and their contents as well as a set of open file format evaluation criteria for assessing a format's "archival ability" and "retrieve ability".

ACKNOWLEDGEMENTS

Fuji Heavy Industries Ltd.	Tetsuo Tominaga
Hino Computer System Co.,Ltd.	Hiroshi Ohta
Honda Motor Co., Ltd.	Koichiro Kawakami
Isuzu Motors Limited	Yoichi Ishida
Kawasaki Heavy Industries, Ltd.	Shinya Sato
Mitsubishi Motors Corporation	Kenji Ando
Mitsubishi Fuso Truck and Bus Corporation	Kunihiko Yoshino
Nissan Motor Co., Ltd.	Masaya Ozawa
Suzuki Motor Corporation	Tetsumi Kobayashi
Toyota Motor Corporation	Tadashi Yamada
Yamaha Motor Co., Ltd.	Takanori Toguchi
Digital Process Ltd.	Takamasa Tanaka
Fujitsu Kyushu Systems Limited	Mari Omura
Fujitsu Kyushu Systems Limited	Katsuya Toyama
PSA Peugeot Citroën	Frédéric Chambolle
International TechneGroup Incorporated	Mike Lemon
Cape Cod Community College	Fredrick Bsharah

TABLE OF CONTENTS

Figures

Tables

1 INTRODUCTION: LONG TERM ARCHIVING& RETRIEVAL (LTAR)

In today's engineering and manufacturing organizations, paper based product design and analysis approaches have been or soon will be replaced by computer-based solutions that digitally store and manage the product definition information. New business processes, information architectures and models, and hardware/software infrastructures have been deployed within the OEM and supply communities to effectively leverage the initial usage of this newly created digital information.

However, the processes, models, and infrastructural designs for addressing the Long Term Archival and reuse of the digital information have not been widely deployed. Long Term archival and reuse has been a challenge because any solution requires alignment of storage media, data architecture, authoring/editing software, and hardware infrastructure. Such an alignment can be difficult to achieve because each of these components have their own unique lifecycle durations.

Until recently, the relative newness of digitally managed product definition and lifecycle information has afforded companies with the opportunity to ignore Long Term archival issues. However, many companies have now reached a level of maturity with digital product lifecycle information management so that issues pertaining to data retention and reuse have become paramount with respect to their near-term business plans and economic viability.

The recommendations developed by this project have been designed to guide companies to effective and efficient archival and retrieval practices. Specific recommendations address Format, the LTAR Process, Retention Time Period, and Quality Assurance. In addition the project will develop a test bed capability for assessing an enterprise's LTAR capability.

The choice of a LTAR format must be based on clearly defined criteria. This document begins with this focus. Specifically the criteria should take into account the time period that may exist between the archival and the retrieval as well as the possibility that an exchange can occur between 2 different software applications. Namely that the "creating" software and the "reading" (CAD viewer) software may be different and that the LTAR process is typically spans a long time frame.

After the criteria are defined, the document presents a LTAR classification scheme that is partitioned into the three categories of Native, Neutral, and Visualization. This is followed by a description of archival file format risk management. With the classification and risks explained, an assessment of the information types most commonly implemented by the commercial software vendors is documented. This is done from the perspective of the Native, Neutral, and Visualization classification categories. The document concludes with an appendix listing the file formats that were explicitly considered in the development of the recommendation.

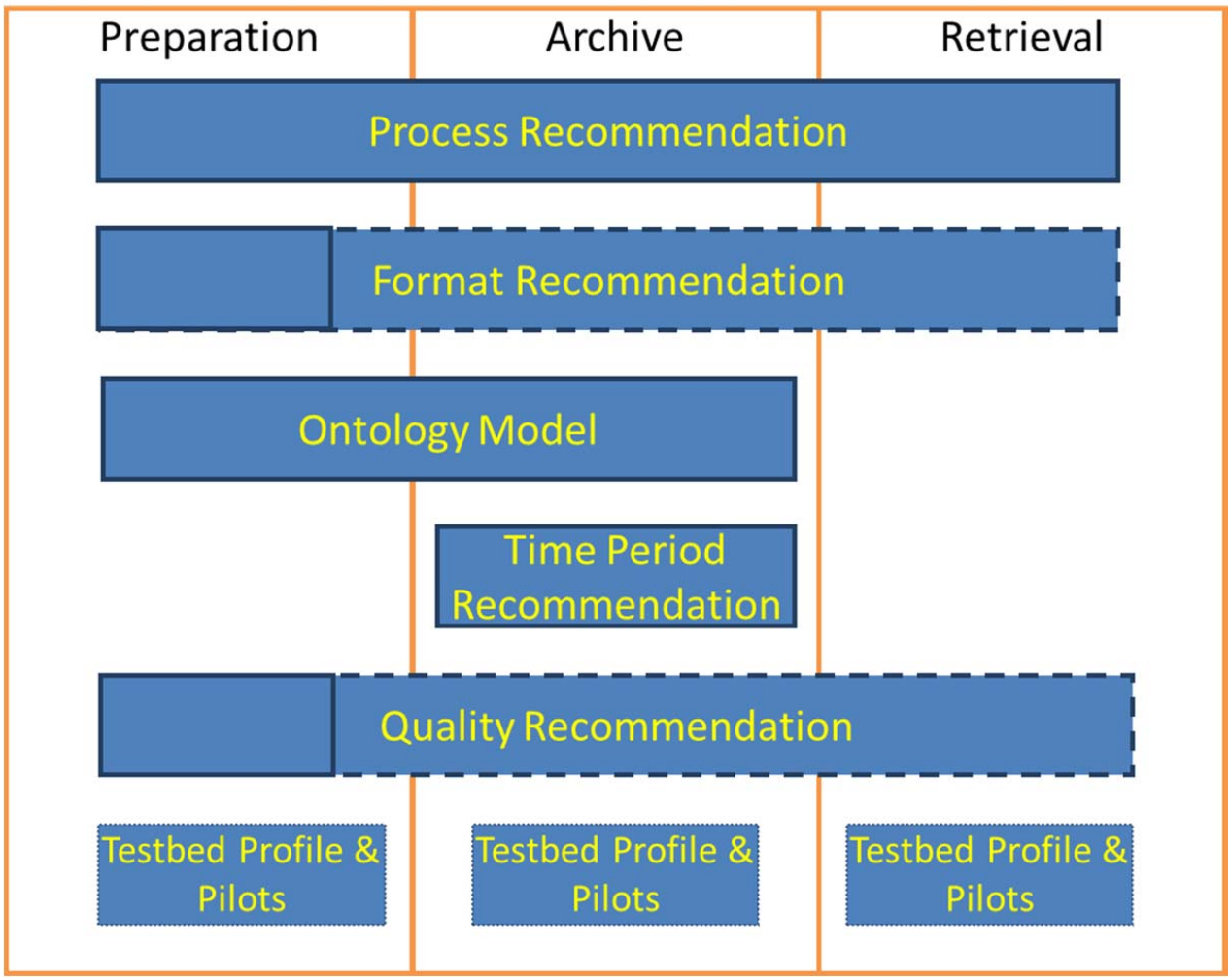


Figure 1: Long Term Archiving Areas of Recommendation

2 LTAR FORMAT RECOMMENDATIONS

2.1 Overview

The purpose of this recommendation document is to provide:

- The criteria for assessing an archival format.
- Classifications for a better understanding of the different types of formats and their contents.

This aims to provide a company key elements to choose its archival format(s) based on functional and technical requirements.

A first list of formats is proposed in the appendix, giving an overview on available formats (native, neutral, visualization) mostly used in the automotive industry. Nevertheless, a company is still free to choose any format it wants to apply.

2.2 Criteria for a LTAR format

The choice of a LTAR format must be based on clearly defined criteria. Those criteria should take into account the LTAR context:

- Time Frame: A long time period may exist between the archiving and the retrieval: During this period, the software product and/or editor of the software used to create the input data may disappear along with its format(s).
- Creating and Reading Software: Exchange can occur between 2 different software products: the fact to archive a data and read far later implicitly means data exchange between the "creating" software product and the "reading" (CAD viewer) software product.

Using an open file format is the SASIG-LTAR recommended direction to support the LTAR context.

An open file format is a published specification for storing digital data, usually maintained by a standards organization. Therefore, the mode of presentation of its data is transparent and/or its specification is publicly available, and thus, implementable by anyone. There are nevertheless cases of open file formats promoted by software companies which choose to make the specification of the formats used by their products publicly available. Open file formats are usually designed to facilitate interoperability between software.

The SASIG-LTAR work group has identified the following open file format evaluation criteria for use in this recommendation. Nevertheless, a company or organization is still free to develop its own LTAR format, but it needs to take into account that the following criteria are satisfied within its user community to guaranty the "archival ability" and "retrieve ability".

Available	Described precisely in a way that is widely and freely available, does not use proprietary methods or practices; (e.g., object modeling methods using UML or EXPRESS), and is exhaustively defined.
Fully Defined	Format and services implementing the data are explicitly described (e.g., STEP Part 21 or XML, PLM services, binary/text formats, etc.) and documentation of the format and services is readily available.

Widely Used	Selecting a file format for its archival strategy requires that the concerned format is widely used and recognized within the corresponding community of users. This criterion should contribute to a better adoption of the archival principle and belonging formats.
Maintainable	The maintaining process is described and well accepted by the applicable community and all organizations in that community are able to comment and participate in the approval process; (e.g., STEP ISO ballot procedures, OMG and W3C consortium procedures). Responsibility for maintaining the standard is clearly defined and held by a responsive organization;
Non-restricted	The standard and its documentation is publicly available, is not restricted by royalties, patents or other Intellectual Property restrictions, except possibly copyright, in which case copies must be available at reasonable cost.

2.3 Format classification

2.3.1 Native format

A native format is always associated to authoring software, and is the unique and proper format used by this software (example: *.DOC* file for Microsoft Word, *.CATPart file* for CATIA V5, *.DWG file* for AutoCAD)**.** It is the format in which authoring software manages the data**.** Thus, it collects the whole richness of the data created using the authoring software functions (example: construction history and feature parameters). The native format contains the complete and original representation of a component, complex or not and it should remain the reference definition but it usually can only be read by the authoring software which was used to create it (or the associated API).

Experience shows that along the lifecycle of the data some limitations may appear through the usage of native formats:

- The native format resulting from translation software cannot be considered as a pure native format. A translation usually cannot conserve the initial richness of the input model.
- Loss of data may appear while reading a given native format with a previous version or release of the corresponding authoring software.

2.3.2 Neutral format

A neutral format is designed to enable exchange of data between different software products, and so called "pivot format". Thus, a neutral format implements its own data model (neutral data model) representing objects and relationships most commonly used by the commercial software products. Consequently, some constructs are not supported by neutral formats because:

- The constructs are too specific, and then impossible to rebuild in the target system (example: feature option that is only used by one system).
- It is difficult to rebuild an identical representation in the target system. In the CAD domain, it is typically the case of the feature-based modeling for which the B-REP

definition depends on the CAD algorithms that implement their own resolution algorithms of singularities.

Therefore, a neutral format is less rich than a native format, and an information loss may occur during the translation process.

The reading and writing of a neutral format from any software product is done through a translation process between the software native format and the neutral format. This translation is based on a mapping table establishing a correspondence between native and neutral object(s). A neutral format is generally generated from the original native format.

Most of the neutral formats are national (JIS, DIN, ANSI, AFNOR), international (ISO), or de-facto standards.

2.3.3 Visualization Format

A visualization format is designed to facilitate the visual manipulation of complex models using relatively "easy-to-use" presentation software product that doesn't usually require having full access to the complexity of the definition. A visualization format is always generated from a native or neutral format, through a translation process. The expected result file is characterized by a smaller file size compared to the original native format file size.

A visualization format shall contain a simplified representation of the initial native format, which contents are aligned with the business requirements. Thus, the result will consist in a reduced and "dead representation"[1]. In the CAD domain, the 3D complex model definition is simplified through a tessellation process that approximates the 3D model with facets. This implies that some operations on a visualization format may be impossible, or lead to an approximate result. The original, exact representation cannot be retrieved from the visualization format; full reverse engineering is not possible.

2.3.4 Relationships and comparisons between formats

The following figure intends to illustrate the logical dependence of native, neutral and visualization format in terms of origin.

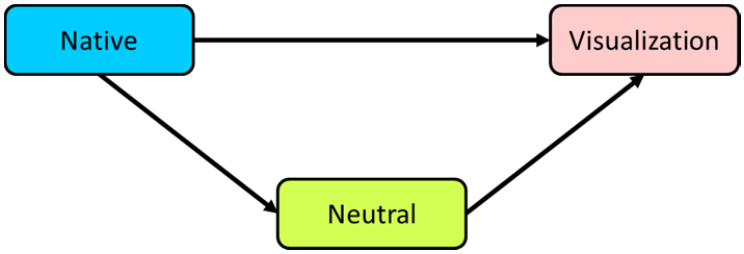


Figure 2: Relationship of Formats

[1] A "dead representation" contains the visualization representation of a3D model, but cannot be easily modified because the original shape definition parameters are not available.

The limitations detailed in section 2.3.1-Native Format show that native formats are not typically suitable for archival purpose. Figure 3 illustrates how the usage of neutral or visualization formats may prevent data loss along the lifecycle of the archived data.

Some file formats may contain visualization and exact representation. In that case, the file size characteristics of visualization format are not fulfilled. If it is used by visualization software, only the visualization representation will be used.

The image below illustrates how using a native archival format versus neutral or visualization archival format may affect the final result when the native software product undergoes version evolution over its life cycle. The lowercase delta (δ) and the capital delta (Δ) represent the data loss. Indeed, if a company selects a native format as archival format, this implies multiple migrations to assure that the associated software will be able to read the retrieved data. Those migrations may introduce data loss. Even if the translation from a native format to an archival format (neutral or visualization format) introduces data loss, this data loss is usually less than the cumulative losses for the native format.

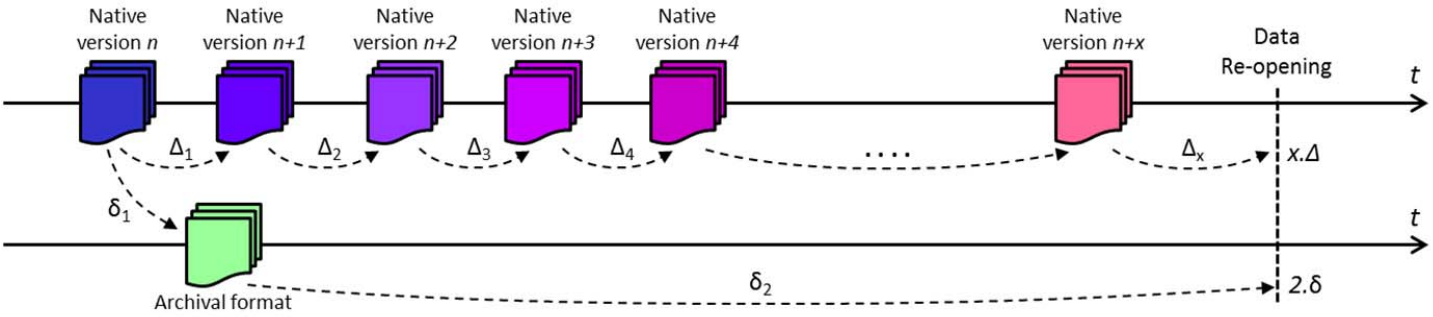


Figure 3: Data Loss Comparison

2.4 Risk management related to archival file formats

2.4.1 Archival to Retrieval Route definition

LTAR can be considered as a long term data exchange. In a typical scenario, data produced by software product "A" gets identified for archival. The data gets archived which may require that it be translated into a specific archival format or that it is archived in the original native format. After many years, software product "B" is used to re-open the data. In that case, an Archival to Retrieval Route long term data exchange is used.

The figure below shows that data exchange from authoring software to reading software along the LTAR Time Period has to pass through some intermediate/relay points. These intermediate or relay points are key elements that guarantee a correct operating LTAR solution.

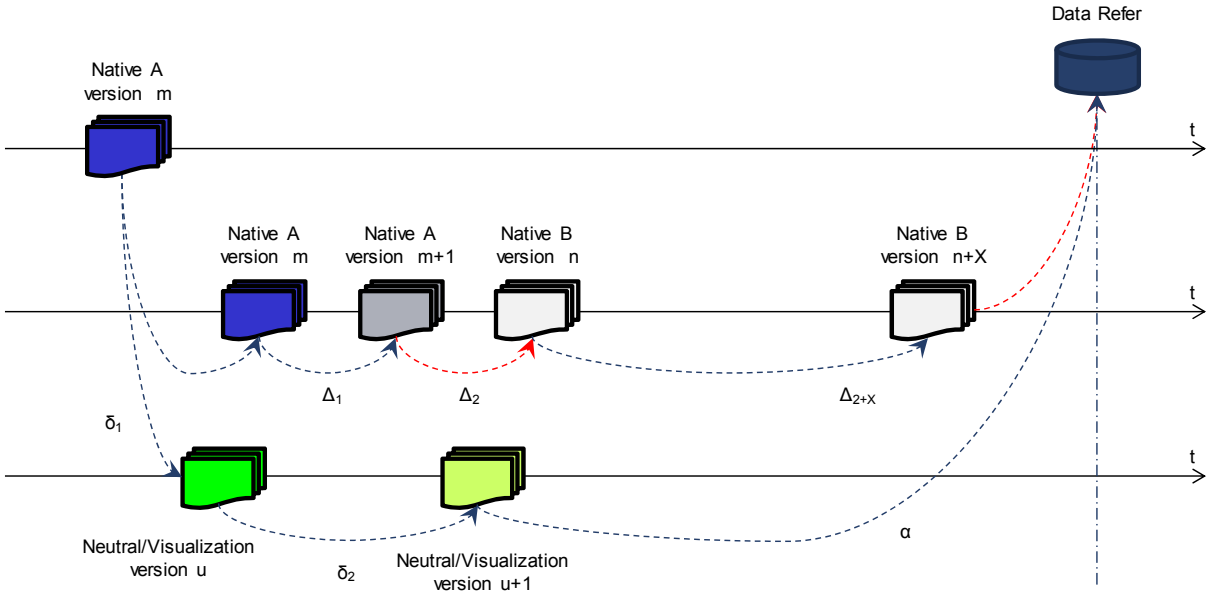
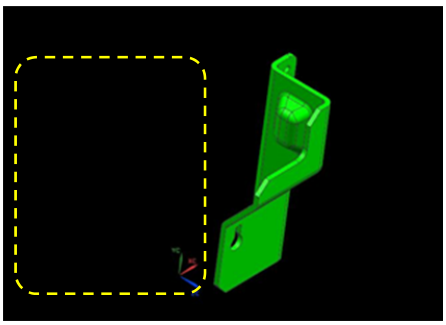


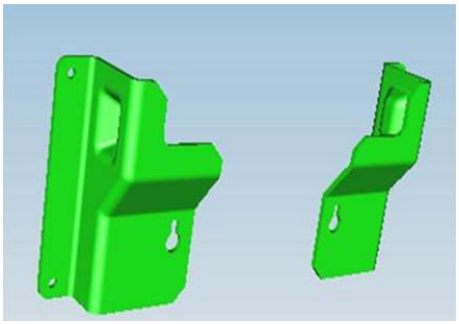
Figure 4: Relay Points in LTAR Time Period

Risks may appear along the Archival to Retrieval Route. These risks can be divided into two sub-categories:

- Continuity Risks: As archived data has to be readable at any time during LTAR time period, the Archival to Retrieval Route must always exist (i.e., not be disconnected). The continuity risk characterizes the need to ensure the format evaluation criteria are continuously satisfied. If not, the Archival to Retrieval Route is disconnected.
- Reproducibility Risks: The expected result of the retrieval process is to recover the archived data as it was before it was archived. Even if the contents of the original data are fully preserved from one perspective, there is still the possibility that associated functionality will be limited with respect to another perspective such as the perspective of the reading (CAD viewer) software product. For example, if the archive data contains the finished representation as well as the hidden construction reference data used to create it, then the support of show/hide setting is expected in the reading software product. That is it is expected to support the display the construction reference data based upon the show/hide variable setting. If this setting is not properly supported, then both data could be displayed which will introduce ambiguity and the Archival to Retrieval Route will be invalid.



Display when archived
(Construction reference data is set as hidden)



Display when retrieval
(Construction reference data is displayed)

Figure 5: Comparison of Reproducibility

2.4.2 Continuity related risks

Continuity related risks pertain to a disconnection in the exchange route. Two main disconnection reasons can occur:

1. Software product is not available for reading the archived data during the time retention period. This situation is due to the lack of support for the archived formats. Therefore, any organization implementing a LTAR project needs to ensure that the software products deployed during the LTAR time period are compliant with their LTAR strategy in terms of archival file format.

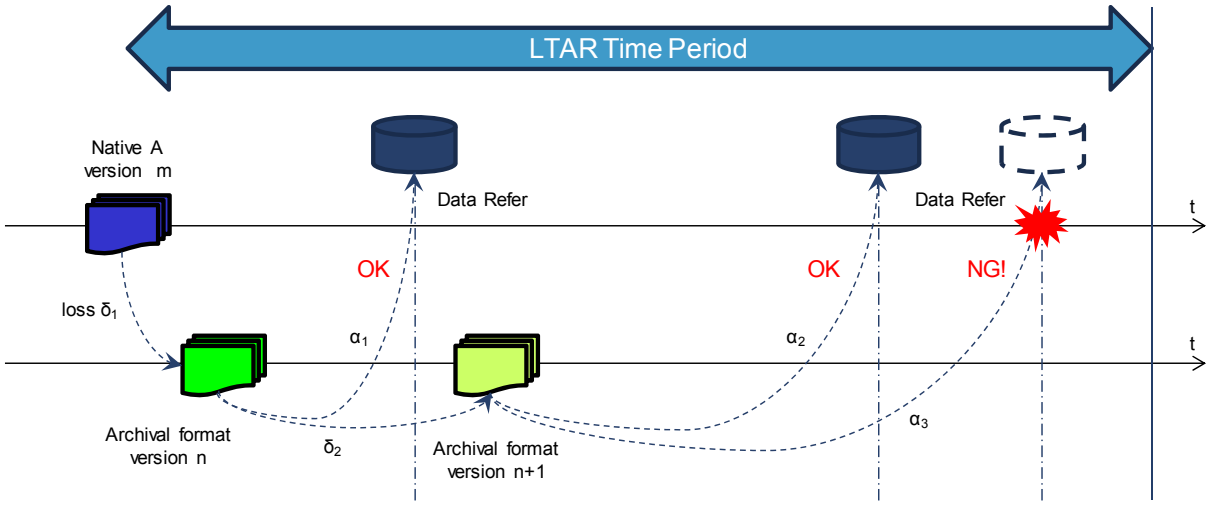


Figure 6: Ensure Software Product Compliancy

To prevent such a situation:

– The organization needs to engage with their software product vendor to ensure support for the archival file format(s).

– The organization needs to select a new archival file format. This step includes a migration of the archived files from the old to the new archival format.

2. The data cannot be properly migrated during LTAR time period. This situation often arises when a native file format LTAR strategy does not support upward compatibility. Upward compatibility of formats refers to the ability of an archive format version n+1 to fully encompass concepts supported under version n. By this definition, if previous versions (n, n-1, n-2, etcetera) were also upward compatible, then, by induction, version n+1 will accept input that worked under any prior version. However when a supported archival format is deprecated (obsolete or no longer supported) and not supported in newer releases of the archive format through either direct mapping or translation then a continuity disconnect results.

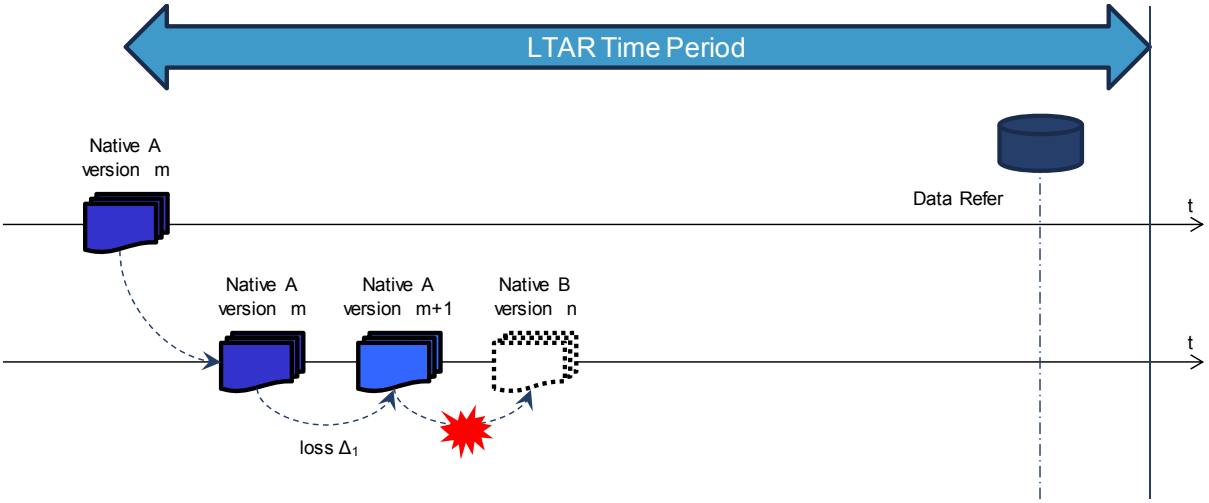


Figure 7: Upward Compatibility in Native Format

To prevent such a situation:

– The organization needs to engage with their software product vendor to implement better quality translation software.

– The organization should migrate the native format into a neutral or visualization archival format. The selection of either a neutral or visualization archival format is dependent on the business requirements and the archived data contents (exact versus visualization).

2.4.3 Reproducibility related risks

Reproducibility risks relate to data loss during exchange. Two main reasons for data exchange loss may occur:

1. Too many migrations can cause serious data loss. Depending on the format strategy and business requirements, an organization may have to migrate many times during the LTAR time period the data archived. Each migration may introduce data loss; thus the cumulative loss of fidelity can problems associated with the reading of the archived data, which in turn impacts business activities.

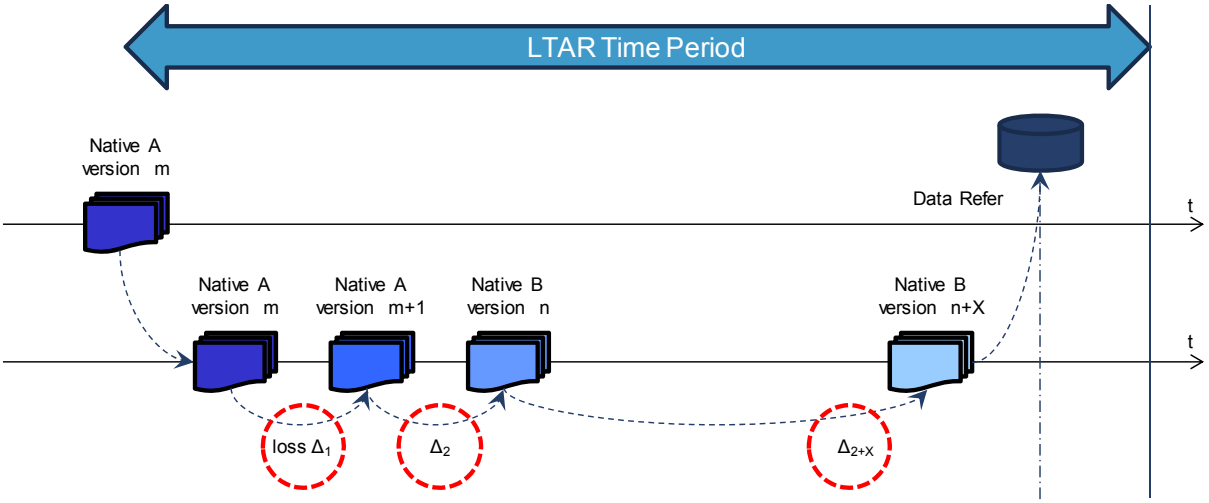


Figure 8: Migration Data Loss

To prevent such a situation:

- The organization should reduce as much as possible the number of migrations.
- The organization needs to engage with their software product vendor to reduce the migration data loss, or to provide the best level of compatibility as possible between the old and the new archive formats.

2. The reading software product (convertor, viewer, etcetera) doesn't provide an acceptable level of support of the archival format. The reading software product used by the organization doesn't provide enough functionality to read the archive format with a level of translation knowledge that is acceptable with respect to the business's retrieval requirements.

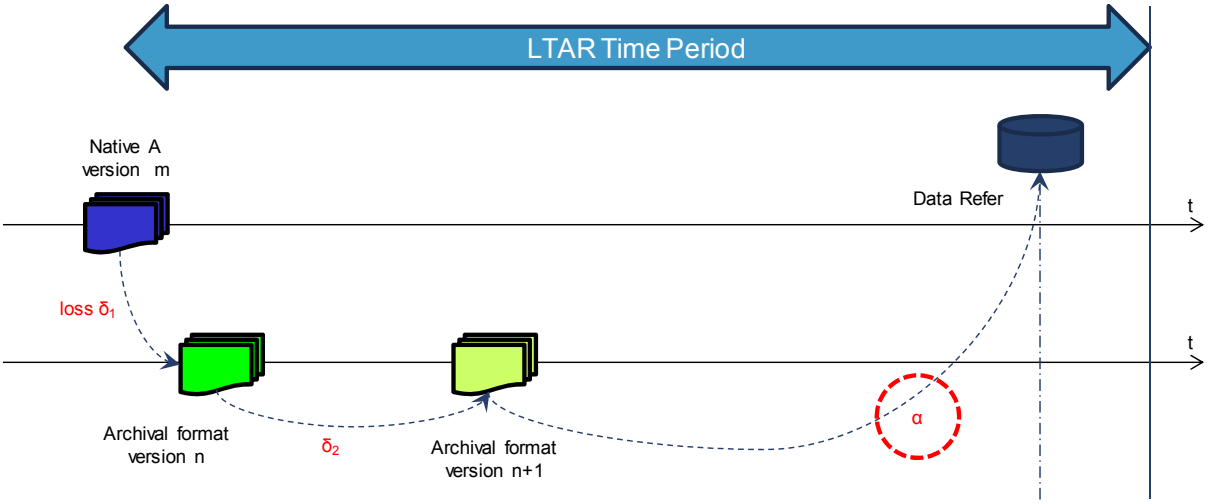


Figure 9: Reading Software Unacceptable Level of Support

To prevent such a situation:

- The organization needs to engage with their reading software product vendor to implement better support of the archival format with respect to the business retrieval requirements.
- The organization should migrate its archive data into a better supported archival format. However, the archival format strategy should be independent as much as possible from the software product vendor's product strategy.

2.4.4 Risks assessment

The following table explains the continuity and reproducibility risk assessment for each format. Risk is calculated by multiplying severity and likelihood.

Severity is the consequences of impact to business if the risk occurs.

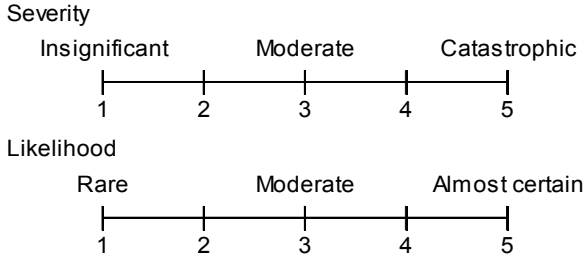
Likelihood is the probability that the risk will occur.

Scales from one to five were used for both severity and likelihood, with a value of five for highest severity or likeliness.

Table 1 Continuity and Reproducibility Risk for Formats.

Category	Description of Risk	Native		Neutral		Visualization	
Continuity	Unable to perform migration e.g. because of significant difference in algorithm between source format and target format.	20		3		8	
		5	4	3	1	4	2
	Support / maintenance of the format is no longer available e.g. because format vendor changed their own policy and quitted to support, or withdrawn from the market.	25		6		16	
		5	5	3	2	4	4
	Reading software which supports to load the format is no longer available (e.g. Reading software vendor withdrawn from the market, or vender quit to provide function to load the format.)	20		4		12	
		5	4	2	2	4	3
Reproducibility	Significant delta causes during one or multiple migration(s). (e.g. Difference in algorithm between source format and target format.)	16		4		9	
		4	4	2	2	3	3
	Reading software shows different result from as it was when it was archived because of different interpretation between software. (Interpretation to dialect, software specific extension/customization of the format.)	4		8		6	
		4	1	2	4	3	2
	Upgraded reading software shows different result because of changes in functions or algorithm of the software.	12		6		9	
		4	3	2	3	3	3

Scales



Risk Rating (S x L)	
Severity	Likelihood

2.5 Domain coverage by format type

This section provides a document domain coverage mapping for the three file types of native, neutral and visualization. The document domain areas include office documents, 2D & graphic documents, and 3D documents. Each of the three document domain areas are decomposed into subject areas that reflect the type of information that is most commonly implemented by the commercial software product vendors. This evaluation is based on what can be saved with a particular format, not what can be reproduced.

This evaluation corresponds to the state of the art when this document was written. However, it is possible that this evaluation will change in the future due to format enhancements.

In the following tables, the visual assessment used should be interpreted as follows:

- Full circle: Full support.
- Three-quarters, One-half, One-quarter: Partial support. More the number of quarters, better is the level of support.
- Empty circle: No support.

2.5.1 Office documents' family

Content	Native	Neutral	Visualization
Formatted Text	●	◕	◡ (1)
Graphics	●	◕	◡ (1)
Tables	●	◡ (1)	◡ (1)
Formula	●	◡ (1)	○

Table 2 : Domain covering for Office documents

(1) : Only the presentation of the resulting data is available.

2.5.2 2D & Graphic documents' family

Content	Native	Neutral	Visualization
2D Wireframe representation	●	●	◡ (1)
Color & Texture	●	◡	◡ (1)
Layer structure	●	◡	◡
2D Annotation	●	◡	◡ (1)
2D Dimensioning &Tolerancing	●	◡	◡ (1)
Graphical image	●	●	●

Table 3 : Domain covering for 2D documents

(1) : Only the presentation of the resulting data is available.

2.5.3 3D documents' family

Content	Native	Neutral	Visualization
3D Exact representation	●	●	◗ (1)
3D Facetted representation	●	◡	●
Construction history	●	◡	○
Features	●	◡	○
Knowledge	●	◗	○
3D Dimensioning & Tolerancing	●	◔ (2)	◡ (3)
3D Annotations	●	◔ (2)	◡ (3)
Graphical properties (color, lighting,…)	●	●	◔
Assembly (Product Structure)	●	●	◔
Product information	●	●	◗ (3)
Validation property	○	●	◡ (3)

Table 4 : Domain covering for 3D formats

(1) : Certain lightweight formats may contain some 3D exact elements in addition to the 3D facetted representation.

(2) : Certain neutral formats may contain the structured definition with the associated presentation, and some other ones, only the presentation.

(3) : Visual presentation only available. In few cases, representation is also available.

3 MAINTENANCE REQUEST

If you find an error or other changes that should be made to this publication, please complete this form and return it to the proper address below.

Name of Submitter: Date:
Company:
Company Address:
Phone: Fax: E-mail:
CHANGE REQUEST (Use additional sheets if necessary)
Page Number of Change:
Document Currently Reads:

Recommended Changes/Should Read:

Recommended Additions:

Reason for Change:

Signature of Submitter:

Submit This Change Request to Your Organization at the Address Listed Below:

Automotive Industry Action Group
26200 Lahser Road, Suite 200
Southfield, MI 48034
USA
Phone: (248) 358-3570
Fax: (248) 358-3253
Web: www.aiag.org

Groupement pour l'Amélioration des Liaisons dans l'Industrie Automobile
20, rue Danjou
92100 Boulogne Billancourt
France
Phone: +33 1 41 31 68 68
Fax: +33 1 41 31 68 60
Web: www.galia.com

Japan Automobile Manufacturers Association, Inc. (JAMA)
Jidosha Kaikan,
1-30, Shiba Daimon 1-chome, Minato-ku,
Tokyo 105-0012
Japan
Phone: +81-3-5405-6130
Fax: +81-3-5405-6136
Web: www.jama.or.jp