解 説 4

1,000年アーカイブ
― デジタル情報長期保存の考え方 ―

パソコンをはじめ、デジタルカメラ、スマートフォンや携帯電話などの携帯端末、ビデオ・レコーダーなどさまざまな機器によってデジタル情報が生成・保存されています。これらのデジタル化された情報は、従来の紙やフィルムなどの情報とは異なった方法での保存が必要となっています。

今でも、大掃除や引っ越しなどの機会に、はるか昔の手紙や写真などが見つかって、あまりの懐かしさに作業する手がしばらく止まってしまうことがあります。また、新聞などを読んでいても、何百年も前の古文書が発見され、その当時を研究する上での重要な資料となるといった記事を目にしたりします。これがデジタル化された情報であった場合、同じことが起きるでしょうか。先日オフィスの引っ越しの際に、引き出しの奥から 5.25 インチのフロッピー・ディスクが何枚か出てきました。張られているラベルを見ると、25 年ほど前に書いたレポートが保存されているようですが、どうにもそれを読み出して元の文書に戻すすべがありません。結局ゴミ箱行きになりました。

この違いはどこから来るのでしょうか。本技術解説では、これまで長期に保存されてきた情報の種類と長期保存実現の背景、デジタル情報としての長期保存の考え方と最近の IT 業界での取り組み例を解説し、例えば 1,000 年にも及ぶデジタル情報長期保存（1,000 年アーカイブ）について考えたいと思います。

❶ 今までの長期情報保存

今まで残されてきた従来媒体（非 IT 媒体）の情報は、原本がそのまま保存されてきたものと、コピーを繰り返して残されてきたものの 2 タイプに分けて考えることができます。前者を物理保存（Physical Preservation）、後者を論理保存（Logical Preservation）と呼びます。ここでは、それぞれのタイプについて解説します。

1.1 物理保存（Physical Preservation）

出土した木簡に最古の平仮名が書かれていたという記事を目にしたことがあります。物理保存の対象はこの木簡のような媒体で、原本そのままを保存することが基本となります。木簡によっては書かれている文字がなかなか識別できない場合もあるでしょう。ロゼッタ・ストーンなどは、発見当時は書かれている文字が読めなかったそうです。このことから、その残されている記号・文字自体（後の研究で文字でなかったと分かるかもしれませんが）も情報として考古学的に高い価値があり、その記号・文字が伝えようとする内容がその価値をさらに高めていることが分かります。このような保存を IT 的に考えた場合をビット保存（Bit Preservation）と呼びます。ビット列さえ残っていれば、少なくとも信号として取り出せる可能性があり、そのビット列が伝えようとする内容を復元する技術があるという前提での保存といえます。しかし個々の技術や製品のライフサイクルをはるかに超える期間、この前提を成立・維持させるのは容易ではありません。

1.2 論理保存（Logical Preservation）

源氏物語は 1,000 年ほど前に書かれた世界最古の小説といわれていますが、今でも小説や映画、ドラマ、コミック本などに再編集され、多くの人々に親しまれています。一方、この原本の存在ははっきりしないようです。このタイプの保存は、原本に書かれていた内容を繰り返しコピーすることによって実現しています。場合によっては、部分的に再編集が加えられたかもしれませんが、光源氏を中心としたストーリーであることは一貫しています。1,000 年の間、繰り返し行われたコピーによって保存された情報こそがこのストーリー、すなわち情報が持つ意味（内容）だったわけです。このような保存を論理保存と呼んでいます。これを IT 的に考えた場合、必要に応じてデータのフォーマット変更を伴いながら、その都度最新の技術を使って進めるデータ・マイグレーション・サイクルの実行としてとらえることができます。

1.3 保存目的

仮に原本を大切に保存するとしても手間が掛かります。ましてや、コピー（写本）を長期にわたり繰り返すとなる

となおさら大変です。源氏物語はその内容が多くの人に支持された（ニーズがあった）ことから、時代を問わず多くの人が手元に置くために、手間を掛けてコピーを繰り返してきたと考えることができます。文字以外の情報についても、繰り返しコピーを行うことによって伝承されるものがあります。ちょうど今年は伊勢神宮の式年遷宮の年に当たります。20 年周期でお社を建て替えるもので、この繰り返しにより建築技術と儀式の保存が行われてきました。この儀式の持つ特別な意味が、式年遷宮をおよそ 1,300 年の間継続させてきたといえます。見方を変えれば、掛ける手間（コスト）に見合うだけの価値を見いだせなければ、論理保存（データ・マイグレーションの実行）が行われることは非常に難しい（消えてなくなる）と見ることもできます。物理保存では、ロゼッタ・ストーンや押し入れの奥から出てきた古い写真のように、特別にコストを掛けなくても残せるものが結果的に残ったといえるかもしれません。

さて、デジタル情報に目を向けた場合、その長期保存はどのような観点で検討しなければならないのでしょうか。

❷ IT におけるアーカイブの取り組み

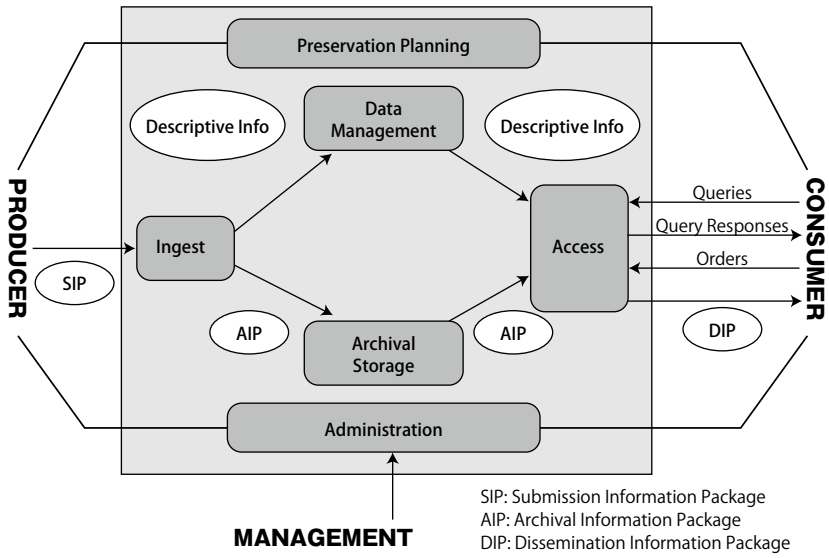
2.1 アーカイブのモデル化

アーカイブに限りませんが、対象となる要件をシステムとして考える場合、そのモデル化から始めるのが一般的です。モデル化によって、一体何をするシステムなのか、入力は何で、どのような処理を内部で行い、何を出力するのかといったシステムの役割や構成の定義に加え、システムを議論する上での用語を関係者で共有すること（同じ言語で議論すること）が可能になります。システムとしてのアーカイブの議論において、おそらく最も多く参照されているのが OAIS（Open Archival Information System）が定義する Reference Model［1］ではないでしょうか。このモデルは**図 1** の要素によって構成されています。このモデルで最も重要な点はパッケージの考え方を導入したことです。ここでのパッケージは**図 2** の要素によって構成されます。この定義は、実際にデジタル化された情報を保存し長期にわたって管理する際に、主たる保存対象（例えば映像ファイル）に加えて、どういった追加情報が必要になるかを教えてくれます。また、このパッケージを、**図 1** のモデル上に置いてみると、それらのパッケージ内要素がどういった局面で参照されるのかが見えてきます。保存期間が長ければ、IT システムの更新が繰り返されることもあるでしょうし、場合によっては管理母体や企業体の間でのシステム移管ということもあるかもしれません。そのようなケースを OAIS モデル上で考察すると、それは OAIS モデル間のパッケージ移行（パッケージ・マイグレーション）ととらえることができます。その際に、必要な管理情報（パッケージの構成要素）が不足していると、その再構築のために余分な手間（コスト）が掛かる心配があります。コストが掛かり過ぎる場合、移管自体が実施されない（その時点ですべて消える）かもしれません。

OAIS のような共通モデルをベースにシステムを考えることで、将来起こるかもしれないさまざまな局面に対応できるアーカイブ・システムの構築を可能にすることが期待されます。

2.2 外部依存性（View Path）

アーカイブ・モデルをベースにシステム要件や構成要素を検討した後、システムの具体的な設計と構築の段階に入ります。実際のシステムはさまざまなハードウェアやソフトウェアの構成コンポーネントによって成り立っています。これらの構成コンポーネントはどれも寿命を持っています。寿命というと、機器が壊れるまでの期間ととらえる場合が少なくないかもしれませんが、システム上の構成コンポーネントとしての寿命は保守期限と考えた方がいい場合が少



出典：REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS) CCSDS 650.0-M-2 [1]

図 1. OAIS Functional Entities

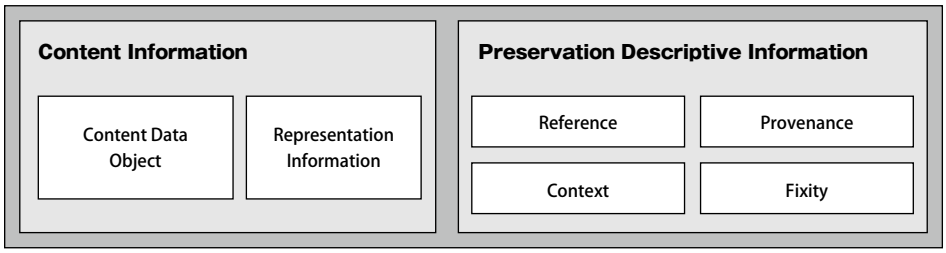


図 2. Structure of Archival Information Package［1］

なくありません。例えば、一部の OS などでは保守期限日以降はウイルス対策などのパッチ提供がなくなるので、通常は問題なく動作していたとしてもセキュリティー上使えなくなったりします。また、各コンポーネントの提供ベンダーが異なると、コンポーネントごとに保守契約が結ばれるので、すべて異なる寿命を持つことになります。

システム・アーキテクチャーをシステム階層で考えることは一般的な手法です。これらのコンポーネントをシステム・アーキテクチャー的に見てみると、それぞれがシステム階層のどれかの層に属していることになります。この観点からアーカイブ・システムを考えれば、アーカイブ・システム全体が動いているということは、目的とするファイル・アクセスのためにすべての階層が機能している（寿命が尽きていない）ということになります。

IBM はオランダ国立図書館と共同で、この観点でのアーカイブ・システム管理を検討しました。検討レポート［2］のキーワードの 1 つは View Path です（**図 3**）。システム階層の各コンポーネントの中には互換性を持ったほかの選択肢が存在する場合があります。例えば、CD-ROM 上に記録された PDF ファイルを保存しているとします。保存は再利用（閲覧）することを目的で行っているわけですから、その PDF ファイルがスクリーン上に表示される必要があります。この CD-ROM とスクリーン上の PDF ファイル表示の間のシステム階層を見た場合、例えば OS であれば Windows 7 と Windows 2008 Server は、対象とするアーカイブ・システム要件上は互換性があるかもしれません。保存されている PDF ファイルについては、対象となる PDF バージョンに対応する Windows 版の PDF ビューアーは幾つも選択肢があるかもしれません。Linux でソフトウェア階層の置き換えも可能かもしれません。ここでの幾つかの選択肢の組み合わせによって構成される CD-ROM からスクリーン上の PDF ファイル表示までのパスを View Path と呼びます。この View Path の数が多いほど、アーカイブ・システムはより多くの代替手段を持つことになるので外部依存性が少ない（安心して使える）ということになります。逆に View Path が 1 つしかない保存ファイルは、早急に手を打たないと再利用できなくなるかもしれません。つまり、長期アーカイブ・システムの管理においては、View Path の観点から各保存ファイルを管理するべきであるという考え方が肝になります。

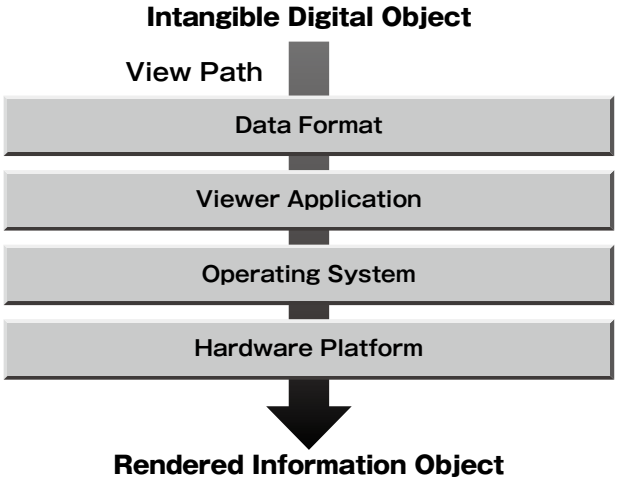


図 3. システム階層と View Path［1］

アプリケーションや OS、それらを実行するハードウェア・プラットフォームや入出力インターフェース、さらにはストレージ・メディアやフォーマット形式などの組み合わせや互換性を考慮すると、実際にはこの考え方に基づいてシステム管理を運用するのはかなりの労力が必要かもしれません。結果的に、View Path に登場したすべての構成コンポーネントを大事に取っておくという博物館的なアプローチを選択することになるかもしれませんが、これはとても大変な作業になります。むしろ積極的にマイグレーションを行って、最新の（これからの寿命が長い）技術で View Path をリフレッシュする作業を選択すべきでしょう。

2.3 アーカイブ・コスト（Archive TCO）

アーカイブ・システム検討で最も悩ましい課題がアーカイブ管理コストです。従来のアーカイブ・メディア、例えば紙、フィルム、ビデオテープなどは、管理コストが余り掛からなかったから残ったといえます。これをデジタル化しファイルとして保存した場合、その管理コストはどう考えたらよいでしょう。一般的に、ただ単に従来メディアをデジタル化しただけでは、その再利用頻度が上がるわけではありま

せん。これを企業活動で考えた場合、ビジネス・モデルなどを変えない限り、個々のアーカイブ・コンテンツが生み出す売り上げ期待値は変わらないことになります。収入は変わらないのに、支出（管理コスト）が増えるということは、単純に利益を圧縮することを意味します。これをデジタル・ジレンマ［3］と呼びます。デジタル化は多くの恩恵があるけれどもビジネスの収支として考えると難しい場合があるということです。

収支バランスを取るためには、1）アーカイブ導入・管理コストを下げる、2）アーカイブからより多くの利益（投資対効果）を得る、の2つの方法が考えられます。アーカイブ導入・管理コストを下げるには、利用するアーカイブ・テクノロジーの選択が重要になります。アーカイブからの利益を上げるには、より積極的にアーカイブ・コンテンツ（商品）をアーカイブ利用者（消費者）に届ける（消費してもらう）努力が必要になります。

❸ アーカイブの収支バランス

3.1 アーカイブ・テクノロジーの選択

View Pathを増やすことがアーカイブ・システム管理上重要であることは先に解説した通りです。これをストレージの立ち位置から見ると、交換メディアの利用、それもオープン・テクノロジーに基づいた交換メディアであることが重要となります。CD-ROMが典型的な例になりますが、ドライブは多くのベンダーから大量に供給されており、ほとんどの商用OSがそのファイル・システムをサポートしています。つまりCD-ROMでのファイル保存は、そのファイルを扱うアプリケーションを除いて、数多くのView Pathを持つことが可能です。また交換メディア自体の管理は棚置きなど、非常に低い管理コストで対応することができます。言い換えれば、伝統的な交換メディアであるフィルムが長期に保存管理できたのであれば、それをデジタル化しIT交換メディアに記録したとしても保存管理できるはずです。

日々大量に生成されているデジタル・データに占めるマルチメディア・ファイルの割合は少なくありません。わたしの家庭ですら、テラバイト・クラスのストレージに大量の映像ファイルが蓄積されています。これらの大量のファイルを効率よく保存するには、大容量の交換メディアが求められます。特にそれらのファイルのコピーを考えると、より高速にデータ転送可能なメディアである必要があります。

この用途に利用できるメディアとしてLTO（Linear Tape Open）テープ（**図4**）が選択肢の1つとして挙げられます［4］。LTOテープはLTOコンソーシアムによってライセンス管理されているオープン・テクノロジーで、2000年から延べ1億5,000万本以上の出荷実績のある磁気テープ・メディアです。最新世代のLTO-6ではテープ1本当たり2.5TBの容量があり、テープ・ドライブ、テープ・メディアともに複数のベンダーが提供しています。データ転送速度もLTO-6の場合160MB/sあり、短時間に大量のデータ読み書きが可能で、現在データセンターや大量のマルチメディア・ファイルを扱う放送局などのアーカイブ・メディアとして世界中で利用が始まっています。

LTOテープは大容量でデータ転送速度も高速なメディアとして有力な候補なのですが、アーカイブ・メディアとして利用検討が始まった6年前には大きな課題がありました。その当時、LTOテープは主としてバックアップ用メディアとして使用されていました。テープ上のファイルへのアクセスにはバックアップ・ソフトウェアなど、専用ソフトウェアの利用が前提となっており、またテープ上のファイル保存フォーマットも専用ソフトウェア間で互換性がありませんでし

出展：LTOコンソーシアムホームページ
http://www.ultrium.com
注：上ホームページは、LTOドライブ圧縮機能利用時の値を掲載しています。映像ファイルはすでに圧縮済みなので、LTOドライブ圧縮機能による再圧縮は期待できません。よって、本資料は非圧縮値で記載しています。

図4. LTOテープとロードマップ

た。これは View Path 確保の観点からも不利です。

そこで IBM はテープ・メディア用のファイル・システムである LTFS（Linear Tape File System）[5] を新たに開発しました（**図 5**）。以前からテープ・メディア用ファイル・システムの研究はあったようですが、なかなか実用には至っていなかったようです。それはテープが追記型のメディアであることに起因します。ディレクトリーを作成したりファイルを保存したりすると、更新されたファイル・システム情報がテープ上に追記されるのですが、結果的にファイル本体とファイル・システム情報がテープ上で入り交じりながら拡散して記録されてしまいます。テープをファイル・システムにマウントする時にはテープからファイル・システム情報（ファイル・リストやディレクトリー情報）を読み出す必要があり、このマウント動作に時間がかかってしまいました。LTO-5 から、新たにマルチパーティション機能がサポートされ、1 本の物理テープを複数の論理テープに分割することが可能になりました。ハードディスク上でドライブ C に加え、パーティションを切ってドライブ D や E を用意するのと同じです。LTFS では**図 5** にあるように LTO テープを 2 つのパーティションに分割し、1 つはデータ・パーティションとしてファイル本体を、もう 1 つ（LTO-5 の場合テープ全体容量の 2.5%）にはインデックス・パーティションとしてファイル・システム情報を保存します。双方を分けて保存すると同時に、ファイル・システム情報の記録をテープ上で局所化することで、短時間でのマウント動作完了を実現しました。

IBM は LTFS の仕様とそのソースコードを公開することで、IBM 以外のベンダーからの LTFS 提供を可能としました。現在、複数のベンダーが LTFS 対応ソフトウェアを出荷しています。これにより LTO をアーカイブ・メディアとして使用する場合の View Path の懸念が払拭されました。

3.2 アーカイブの再利用促進

現在、日々蓄積されていくアーカイブは巨大になります。文書、図面、データなど、大量のファイルが蓄積され、数万時間分のビデオ・ファイルをアーカイブする放送局も珍しくありません。一般にこれらのアーカイブ・コンテンツにはメタ・データ情報が付加されます。これは OAIS のパッケージで紹介した Content Information の Representation Information（技術メタ・データ）と Preservation Description Information（記述メタ・データ）に対応します。技術メタ・データは、ファイル生成環境（アプリケーション種別や準拠するスタンダードなど）や各種パラメーター（解像度など）といった技術的な情報で構成されます。記述メタ・データは、そのファイルの持つ意味的なメタ・データで、図面であれば対象製品名、映像であれば撮影場所や被写体の説明など、観測データであれば観測対象物名や観測条件などがその例となります。現在これらのメタ・データはアーカイブ・コンテンツ登録時に入力されるのが一般的ですが、運用形態によっては入力作業を人手に頼る部分も多く、メタ・データの品質や粒度のばらつきが問題となる場合があります。また、一度に大量のコンテンツをアーカイブする場合は、メタ・データ入力自体が困難になる場合もあります。一方、メタ・データは大量のアーカイブ・コンテンツから目的とするコンテンツを探し出す上で必須になりますので、その品質が良くないと、そのコンテンツが検索でヒットしない（存在が認識されない）という事態を招きます。この課題に対する 1 つの試みとして、動画ファイルについてのメタ・データ入力支援システムが NHK 放送技術研究所によって開発され、

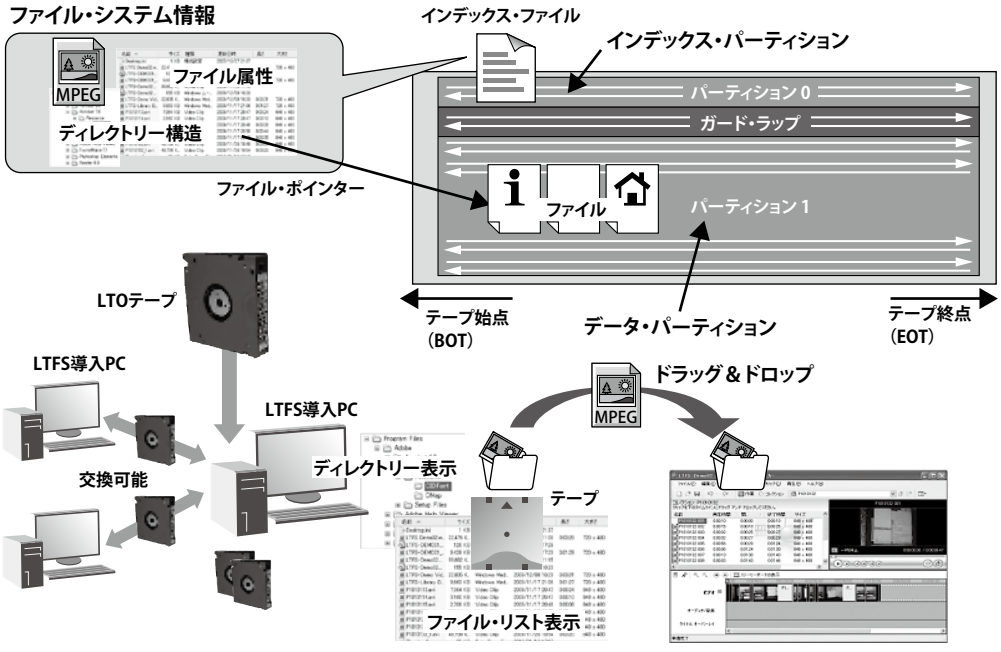


図 5. Linear Tape File System（LTFS）の仕組み

NHK 技研公開 2012 にて展示されました［6］。これは東日本大震災で大量に作られた映像ファイルに対するメタ・データ入力支援をソフトウェアによって行うものです。音声と映像の特徴を時間軸上で抽出し、その組み合わせによって撮影シーンの意味的解釈（例えば、空撮シーンやインタビュー・シーンなど）を行うことでメタ・データ入力を支援します。この入力支援システムは、音声解析と映像解析によるメタ・データ入力のあるべき姿の 1 つを表しています。より詳細な意味的解釈を行うためには、アーカイブ・コンテンツに記録される対象物の判別と対象物の周辺情報（動画や音声の場合は時間的な前後関係情報）との意味的関係をとらえる必要があります。例えば、動画ファイルから「東京駅に到着する新幹線のぞみ号」が映っているシーンを探す場合、駅や列車の識別に加え、出発か到着かを判別します。このような判別や構成要素の組み合わせによって得られる意味的な解釈を行うには、多くの知識が必要です。IBM が開発した質問応答システム Watson がクイズ番組で見せた、自然言語処理などの高い解析技術に今後期待が集まっていくものと思われます。

より詳細なメタ・データ付けによって、より細かな粒度での検索とアーカイブ・プールからの選択的なコンテンツの抽出が可能となります。これを再利用（消費）に結び付けるには、そのコンテンツを求めているユーザーに届ける必要があります。それぞれのユーザーが何をいつ求めるかをユーザーの再利用（消費）動向などによって類推することができれば、ユーザーが求めるコンテンツ候補をタイムリーに提案することができるはずです。すでに実用が始まっているビッグデータ分析技術などにより、ユーザーごとにカスタマイズされたコンテンツ提案が可能となっていくものと期待されます。

❹ おわりに

1,000 年にも及ぶデジタル情報長期保存を考える場合、保存対象物を再利用するための仕組み自体の保存（View Path の確保）と、その対象物を残す動機の明確化が必要になります。オープンでコスト的にも運用可能なテクノロジーと標準的なシステム・アーキテクチャーを利用してその仕組みを実現させることで、将来発生するかもしれないさまざまな変化に柔軟に対応できる可能性が高くなります。

長期にわたるマイグレーション（論理保存）の動機を維持するには、1,000 年もの間残ってきた源氏物語や式年遷宮などの例から、情報の価値の明確化と持続的なニーズの確保について学ぶ必要があります。ただし、巨大なアーカイブ・プールから自分が欲しいと思うものを自分で探すのは至難の業で、保存された情報の存在意義を弱めてしまいます。先進的なテクノロジーを活用することで、ユーザーが求めるであろうアーカイブ・コンテンツ（Right Thing）を、ユーザーに正しい方法で提示（Right Way）し、欲しいと思ったとき（Right Time）に提供することが可能となります。

言い換えれば、この 3 つの R を実現するように先進的なテクノロジー活用を継続的に目指していけば、1,000 年にわたる長期アーカイブが達成できる可能性が高くなります。

［参考文献］

［1］ CCSDS: Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-M-2, http://public.ccsds.org/publications/archive/650x0m2.pdf

［2］ Dr. Raymond J. van Diessen: Preservation Requirements in a Deposit System, IBM/KBLong-Term Preservation Study, http://www.kb.nl/sites/default/files/docs/3-preservation.pdf

［3］ The Academy of Motion Picture Arts and Sciences: The Digital Dillemma, http://www.oscars.org/science-technology/council/projects/digitaldilemma/

［4］ IBM Redbooks: IBM System Storage Tape Library Guide for Open Systems, http://www.redbooks.ibm.com/abstracts/sg245946.html

［5］ IBM Redbooks: IBM Linear Tape File System Installation and Configuration, http://www.redbooks.ibm.com/abstracts/sg248090.html

［6］ 住吉英樹,河合吉彦,望月貴裕,佐野雅規,藤井真人:大震災アーカイブス　メタデータ補完システムの試作,2012年映像情報メディア学会年次大会講演予稿集,6-1,(2012).

日本アイ・ビー・エム株式会社
システム・テクノロジー開発製造.
ストレージ・システムズ開発
シニアエンジニア、
アーカイブシステム・アーキテクト&エバンジェリスト

藤原 忍 **Shinobu Fujihara**

［プロフィール］

1988 年入社。入社以来 SCSI ディスクドライブから始まり、RAID システム、仮想テープシステムおよびストレージ管理インターフェースの開発などに従事。現在は、学会や産業団体、標準化団体活動を通じビッグデータ時代における長期アーカイブ・テクノロジー、特にテープ・テクノロジーの普及とアーカイブ・システム提案に貢献。2011 年情報文化学会賞受賞。IBM Academy of Technology メンバー。