Title	ビッグデータ時代のアーカイブ : その運用と課題
Sub Title	
Author	藤原, 忍(Fujihara, Shinobu)
Publisher	慶應義塾大学デジタルメディア・コンテンツ統合研究センター
Publication year	2014
Jtitle	慶應義塾大学DMC紀要 (DMC Review Keio University). Vol.1, No.1 (2014. 3) ,p.9- 13
JaLC DOI	
Abstract	
Notes	特集 : DMC研究センターシンポジウム : 第3回 デジタル知の文化的普及と深化に向けて : コンテンツとコンテクストの統合的アーカイヴィングに向けて
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO32002001-00000001-0009

ビッグデータ時代のアーカイブーその運用と課題

藤原　忍（日本アイ・ビー・エム株式会社）

日本 IBM の藤原と申します。本日はどうぞよろしくお願いします。今ご紹介にあずかりましたように、私自身は豊洲にある IBM の東京ラボラトリーでアーカイブシステムの開発を担当しております。本日は私どもがここ数年間開発してきましたアーカイブのソリューション、それと今ちょうど三浦さんがご紹介になった内容と、実は非常に重なっている部分がありますビックデータというキーワードをベースにお話をさせていただきたいと思います。幸い私の前に三浦さんが LTO というキーワードを出されました。LTO とはそもそも何なのか、ということになるわけですが、たまたま私の鞄の中に LTO テープが入っていましたので、回覧いたします。どうぞ、みなさんご覧ください。磁気テープになります。

それではお話を始めたいと思います。本日の内容ですが、まずビックデータは以前から別の形で存在していたと。更にその総量は莫大で、実際にそれを蓄積していくということに関しましてはアーカイブテクノロジー、適切なテクノロジーの選択が必要になってくること。更にアーカイブというのは溜めるということを目的にしているわけではなく、これは再利用することが目的になっています。ここから考えましてビックデータのアーカイブの課題、更にゴールということで進めてさせていただきたいと思います。

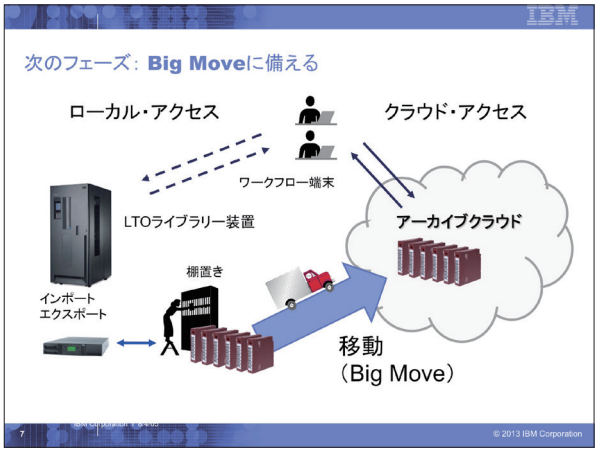
まずビックデータは以前から別の形で存在していたということについてです。これは私自身がいろいろなところへ行きまして撮ってきた写真になっています。私自身がアーカイブということで仕事を始めたのは 2006 年になります。きっかけは NHK の川口にありますアーカイブスセンター。こちらのビデオテープ

100 万本もあるということで、研究所同士の交流会という形で見学に行きました。そのときにわかりましたことは、この膨大なデータが存続の危機にあるということです。というのは、ビデオテープはそこにあるのですけれども、ビデオデッキは機械ですので、保守終了になるわけですね。そうしますと、その 100 万本のビデオテープが再生不可能になるということが、もう 2006 年からわかっていました。実際にいろいろな放送局さんとか回ってみますと、大量のビデオテープがあります。更に大量の DVD があります。ブルーレイがあります。更に大量のフィルムがありますと。これはいわゆる IT 的にみますと IT のデータではなかったのですけれども、これが今 IT のデータにすべて変わろうとしています。実際、どうなっているのか。これは 2008 年当時に、ある放送局さんと試算した結果になっています。その放送局さんには 60 万本のビデオテープがあるということで、その 1 本のビデオテープの尺、さきほど三浦さんからも尺という言葉が出ましたけれども、平均の尺は 45 分でした。ビットレートなども計算しまして、60 万本のビデオテープを MPEG ファイルにファイル変換し、更に新しいカメラが撮ってきた新しいファイル映像を蓄積していくと、9 年後には 15 ペタバイトになるということがわかりました。1 局の話で、です。これが在京キー局さんですとか、中京地区、近畿地区、それから九州とか準キー局さんもいらっしゃいます。そういうところにも、ここまではいかなくてもこれに近い形の、いわゆるデータがあるのです。これをどうするのかということが大きな課題で、さきほどデジタル・ジレンマという言葉が出てきましたけど、あれもそうで、ハリウッドとかインドとか、そういうところで製作されていますデジタルデータを計算してみますと、エクサバイト級になっています。これをどうするかという課題があります。結果的に、当時はケミカルフィルムとして配給されていましたけれども、そのプロセス自身はデジタルですので、基本的にはデジタルデータが作られていくということで。

もう一度繰り返しますと、ビックデータというのは最近の話ではなくて、以前から存在していてアーカイブのテーマだったということです。これまでの議論というのはこの膨大なファイルを何に保存するのかとい

う保存テクノロジーの議論でした。ここにもありますが、真ん中に LTO というふうにあり、磁気テープまたは DVD やブルーレイの光ディスク、または半導体メモリー、HDD、いろいろなものがあるわけです。いわゆる MPEG のようなファイルになったものをいかに保存していくのかということで、保存テクノロジー、ストレージ・テクノロジーを選ばなければいけない。これはどれがいいのかということがここ数年間の大きな議論でした。そのなかで LTO が登場してきました。どうして私の鞄の中に LTO テープが入っているのかというと、LTO がいいだろうと思うから入っているのですが。これまで LTO テープとの比較がいろいろな研究機関で行われてきました。これはハードディスクを作った場合と磁気テープを使った場合のコスト比較ということで行われたものです。上の比較を見ますと、総コスト、つまり初期投資、運用経費、更にフロアの管理費、空調代・電気代全部含めてという意味ですけれども、ハードディスクと磁気テープでは 26 倍のコスト差があります。更にエネルギーでいいますと、105 倍のコスト差があると出てきています。ここから見ますと、おそらくハードディスクの巨大なシステムを構築してそこに溜めるということは現実的ではないですね、ということがわかります。そうしますと、磁気テープのような従来あった媒体ですね、そういうものを活用して保存していくというものがいいだろうなという形の議論になります。システム、設備全体もそういう形の議論になってくる形になります。こういった議論が世界中で行われていまして、私どもメーカーも参加していますし、いろいろな研究機関もこういうようなものを検討されていろいろなリコメンデーションを出しています。これが従来の議論でした。

この議論の先には何があるのでしょうか。ビックデータというのが今日のキーワードになっているわけですが、ビックデータとクラウドというのが非常に密接な関係を持って議論されています。本当に密接な関係があるのかというのはまた別問題かもしれませんけど。今現在におきまして、さきほどのようなテクノロジーを選んで設備を作っていくわけです。最近の言葉でいうとオンプレミスという言葉、プライベートクラウドと言う方もいらっしゃいますが、いわゆる建屋を作りますと。そういう形でコンテンツを溜めこまれていくというのが今の現状です。言い換えますと 100 社あれば、100 社それぞれが投資計画を作って従来のコンテンツをファイルにして溜めこんでいくというのがスタートポイントでした。しかしながらやはりクラウドに移行したいと考える方々が少なくないです。クラウドは、安心して使える、コストが非常に安い、いろいろなメリットがありますので、そちらに移行したいという希望がありますが、ここで大きな問題が起きます。さきほどの 10 ペタバイトだとか 20 ペタバイトという巨大なデータがあった場合に、それをどうやって引っ越すのか、ということです。たとえば巨大なハードディスクみたいなものに溜めていた場合、それをシステムごとデータセンターに送るのかという話になるわけですね。それは基本的に無理です。データセンターはそれを受け取らないです。というのは、ほぼ保守終了になっている枯れた装置をデータセンターは受け取らないです。そうしますと何を送ったらいいでしょうか。たとえば今のビデオテープは？　フィルムは？　と媒体だけ送ってるわけですね。そうすると、ここのシステムからあのシステムに媒体だけ送れるようなからくりを最初から仕込んでおかないと最終的に引っ越しができないということがわかってきました。わかってきたというのは既に設備投資した方のなかでわかってない方は結構いらっしゃるのです。ということは将来的にクラウドに引っ越しできる方と引っ越しできない方に今実は色分けされつつあります。こういうようなことを考えていろいろな方々が設備投資をされています。



たとえば放送局さんで使われているシステムのだいたいの概要ですが、まず編集端末などがあります。アーカイブのサーバーがあって、データベースのサーバーがあります。その後ろのほうを見ますとテープライブラリー装置があって、正副作成と書いてありますのはだいたい 2 本作るんですね。正テープはこのようなオートメーションのなかに、複テープというのは棚置きというような形の管理が非常にポピュラーです。こういうような形のシステムが作られます。問題として上がってきてるのは、ここでキャペックス（CAPEX）、

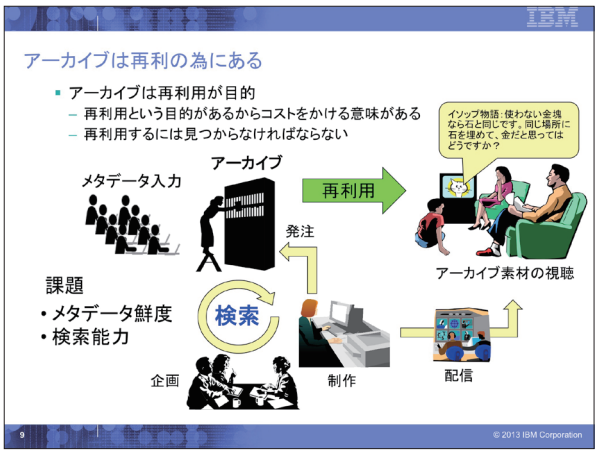
オペックス（OPEX）と書いてありますけど、これはキャピタル・エクスペンディチャー、いわゆる固定資産しての投資ですね。初期投資になります。あと、オペックスというのはオペレーティング・エクスペンディチャー、いわゆる運用経費になります。こういうシステムを導入して、そうすると初期投資としてかなりの金額が必要になります。保守料金ですとか、ここで働く方々の経費も含めてオペレーティング・エクスペンディチャーが発生しますと。問題なのはビデオテープやフィルムをファイル化したからといったとしても、それぞれのコンテンツが稼ぎ出す期待値、いわゆるレベニューには変わらないですね。ビデオテープだから実入りが少ないのかと。これを MPEG ファイルにすると実入りが多くなるのかというと、そういうことは基本的にはないです。たとえば放送局を例にとった場合、チャンネル数は変わっていません。1 日は 24 時間しかありません。そうすると 10 年前に作っていた番組の本数と現在の番組の本数は変わらないのですね。ということは、アーカイブがどんどん増えてコストがどんどん増えていくけれども、使用する量は変わらない。つまりアーカイブの再利用率はどんどん減ってるわけですね。こんなことにお金を使っていていいのかという疑問を持つ経営者の方は少なくないです。これをどうするのだというところに、今たどりついています。ビデオテープからファイルに移行する、フィルムからファイルに移行する。それは映写機がなくなる、ビデオデッキがなくなるという切実な問題がありましたから、勢いでファイル化をしてるわけですが、ファイル化して設備投資した後に、これって何かそんなに正しいことをやってるのだろうか、というところに入っています。

さきほどアーカイブは再利用のためにあるということを申しましたが、これは本当にその通りだと思うのです。ここにも書きましたが、再利用という目的があるから経営上経費、コストをかけることができる、ということがあると思うのですね。ですから、再利用をどんどんしなければいけないわけです。けれども再利用するためには見つからないといけない。何があるかということを見つけなければいけない。さきほど三浦さんの講演でもメタデータという言葉がありましたが、まさにそちらのほうに倒れていく話になります。現在のメタデータ入力というのは人海戦術なんですね。人でしかわからないということで、ビデオを見ながらメタデータを手打ちしています。たとえば NHK さんのようなところでは「映像を見ればだいたい何のことだかわかるからということ」で、退職されて経験が豊富な方々を再採用してメタデータ打ちを頼んでいるそうです。一方、ある在京のキー局では若手の人にやってもらっています。なぜならばコストが安いから。いろんな考え方で人をワーッと並べてやっているという現状があります。それで企画とか制作とか、そういうところで検索をして、再利用して、配信していくと。

イソップ物語に、「使わない金塊なら石と同じです。同じ場所に石を埋めて金だと思ってはどうですか」童話がありますが、これと一緒ですね。使わないと石と変わらないものになってしまいます。

課題として今上がってきたのはメタデータの鮮度と検索能力です。まずメタデータの鮮度というのは、たとえば今日来た、入荷してきたビデオの素材があった場合に、そのメタデータの入力というのは今日の価値観で行うのですね。これに映っているのは何で、どういう内容でというのを打ちますと。そうしますと、メタデータというのは、このコンテンツをアーカイブスに登録したときのメタデータなんですね。10 年前だったら 10 年前のメタデータなのです。それは今の検索する人の価値観と合わない場合があるわけですね。それが 1 つ。いわゆるメタデータの鮮度の問題。もう 1 つは検索能力の問題です。検索能力も、たとえばあるアーカイブスのお話を聞くと、「うちはすごいんだ」と。アーカイブスの棚に行くあの人に「あのときのそれ」と言うとだいたい出てくるというのですね。それでいいのか、という話があるわけです。これは何を言っているかというと、個人の能力に依存している部分が非常に多いということです。見つける側もキーワードを選ばなければいけないので、その選ぶキーワードが非常に個人の能力、経験に依存してしまっています。このことは見つかるものはしょっちゅう見つかるけれど、見つからないものは二度と見つからないということになります。いわゆる検索能力の問題があると。

今どういうところにジャンプしようとしているかというと、機械学習による分析です。人に頼ってはいけないのだというところがあります。まず継続的なメタデータ更新が必要です。鮮度は維持しなければいけない。それから人の能力に依存しない検索手法、手段を提供しなければいけない。もうこうなってくると人では駄目だという話になりますので、機械学習による分析技術になる。機械学習による分析技術って何をしてくれるかというと、特定のコンテンツですね。たとえば IBM のマルチメディア・アナリスティックス & リトリバルシステムで検索した結果を出しているのです



けれど。たとえばアウトドアスポーツ、インドアスポーツ、またはウェディング。そのような抽象的なキーワードで画像を切り出してくるわけですね。切り出すことができるのは、たとえばウェディングだったらどういう映像なのかということをウェディングの状態を撮った写真を 100 枚くらい機械に見せるわけです。それでその映像の特徴を抽出させて学習させますと。フィルターを作りますと。結果的にそのフィルターを使って、今度は大量のデータをバーッと見るわけです。そうすると、一致したものはボーッと出てくるという形になって検出ができるというのがあります。検出ができると、この尺のなかの何時何分何秒にそういうものが映っているのかということが記録できます。これはタグ付けになるわけですね。これでだいたいこの尺のなかにはどういう画像が入ってるのかということがわかるということです。更に、それでも 10 年前には田中投手はいなかったはずですと。今田中投手いますねと。そうすると、「田中投手」ということで検出をさせるわけですね。画像を。10 年前には「田中投手」というキーワードはなかったはずなので、タグ付けもされてないわけですが、今田中投手の顔の写真を見せて、機械に検出させてタグ付けをさせるということをすると。

これだけでいいのか、という話があるわけですが、これだけでは駄目で、次に必要なのはセマンティックなデータへの展開です。意味的解釈。それで非常に残念なことであると同時に非常に嬉しいことでもあるのですが、NHK さんで資料を提供してくださいました。残念なのは、これは IBM の資料ではないということが残念で。でも、NHK さんが提供してくださったということは私たちと NHK さんは仲がいいという意味なのですが。

これは非常におもしろい話で、背景にあったのは 3.11 の震災の映像です。3.11 のときにはカメラマンが東北 3 県に飛び回って大量のビデオを撮りました。数万時間のビデオを撮りました。撮りっぱなしになってしまったのですね。メタデータを付けないと再利用できないという状況になりました。それで NHK 技術研究所である技術を開発しました。これは何を言っているかというと、時間軸で映像が流れるのですが、向こうで見ますと、たとえば撮影メモだとかカラーバーだとか、空撮だとか、顔がある、人の顔があるだとか、いろいろなところがあるわけです。その 1 個の映像を時間軸で見ていって、複数のフィルターをかけているということです。そうすると、いろんなキーワードが出てくるわけですね。時間軸で上から見てみると、意味が解釈できるのではないか、という話です。

次のページに飛びますと、たとえば顔が映っていて人の声がある程度流れている部分。ここはインタビューだとか会見の映像が映ってるのではないか。いわゆる「会見」、または「インタビュー」という解釈になるわけですね。今までは顔だけが検出されました。声だけが検出されましたということで、それぞれの属性が単独で存在していたのですが、それをもう少し有機的につないで意味的な解釈を持たせましょうというところに今発展しようとしています。

そうしますと、田中投手が三振をとってガッツポーズをとった、みたいなところが出てくる可能性があるわけですね。そうすると、たとえばそれで優勝したところだとかなんとかすると、たとえば人の声じゃなくて「歓声」だとか、「ワー！」という歓声だとか、ガッツポーズしている姿だとか、田中投手の顔だとかというのを時間軸で見てみると、そういう瞬間が取り出せますね。それで再利用を上げていきましょうということです。

まとめになりますけれども、機械学習によるメタ付け、タグ付けが必要ですということになるわけです。それはそうですね、という話なのですが、そもそも機械学習できる環境になってるのでしょうか、というところが今度は問題になります。ここで機械による分析が可能な状況で保存されてるということ、というふうに書いてあるのですが、たとえばさきほどの LTO テープも全部棚に置いていたら駄目なんですね。ブルーレイに焼いたとしても、それも棚にあっちゃ駄目だと。要は機械が学習するということはオンラインで、またはオートメーションが容易に検索できる環境じゃないといけないということです。必ずしも LTO テープがすべてオートメーションに入っている必要はないかもしれないです。たとえばその画像かを低解像度、い

わゆるプロキシ画像という形で作って代理検索をさせればいいかもしれないので、必ずしも高解像度のものを全部オンラインにしておく必要はないかもしれないです。いろいろな方法はあると思うのですが、ただし、機械学習が可能な環境をあらかじめ想定してシステムを作っていかないと、持っていても結局 10 年前のメタデータで検索しなければならない。データベースだけは別にあるという形になってしまいます。もう 1 つは仮にオンラインになっていたとしても、それが CPU のメモリーに高速に展開できるものでないといけないと。たとえばDVDにたくさん溜めましたといっても、DVD の呼び出し速度は非常に遅いので、それは CPU の展開に時間がかかるわけですね。そうすると現実問題として機械学習に適さないです。ということで見ると、オンライン可能であり、なおかつ高速に CPU メモリー展開が可能なテクノロジーという形を想定して考えていかないと再利用できるテクノロジーになっていかないだろうと。

まとめの言葉ですが、より高度な再利用が可能なアーカイブのみコストをかける意味がある。リターンがあるからですね。究極的にはアーカイブ自らでコストリカバリーが可能な環境の実現。これが我々の今のゴールになります。これが三浦さんの言われた話とつながっていくというところになるかなというふうに思います。というところで私の部分の話を終わりたいと思います。