失敗に学ぶデジタルアーカイブ
～アーカイブ運営のノウハウを共有する～

岡本　明（NPO法人知的資源イニシアティブ理事、株式会社寿限無代表取締役）

1. はじめに

JPEG(ISO/IEC JTC 1とITU-Tの合同静止画専門委員会)がそのプレスリリースにて国立公文書館のデジタルアーカイブ(以下、「アーカイブ」)を取り上げたのが2004年。それから10年余り経た現在、様々なアーカイブが運営され、多くのコンテンツが公開されて、誰もが手軽に資料を参照できる世の中が到来した。この間、筆者はアーカイブサービスの構築や資料の電子化、関連する規格やガイドラインの整備などに携わる機会を得て比較的多くの事例に触れてきた。

アーカイブの分野にも、充分な評価を得て多くの利用者に参照され更新を重ねていくサービスがある。このようなサービスはマスコミに取り上げられる機会が多く耳目にも触れやすいため、それらがあたかも典型例のように認知されている。

しかしその一方で悩ましい問題を抱えていたり、良い評価を得られずに消えていったりするサービスも少なくない。むしろこちらの方が多く聞こえてくるのだが、世間ではあまり知られていない。

ところが、いざ自分が当事者になろうという段になると、むしろ失敗例のほうが気になるものだ。成功と失敗の事例を比較して勘所を抑えておこうという向きも多い。それなのに肝心の失敗事例が見つけにくい。どこに行けば見られるかと訊かれる。失敗事例が共有されていれば同じ轍を踏むことは無かったのにとの恨み節を耳にする。

そこで、この場を借りて典型的な失敗事例を示すとともに、NPO法人知的資源イニシアティブ(以下、「IRI」)の取り組みを紹介してみたい。なお、失敗例を提示するという本稿の性格上、各事例のサービス名や運営者は一律に匿名とする。ご容赦頂ければ幸いである。

2. アクセスが少ない

ウェブサービスを評価する場合に最も利用される指標はアクセス数だ。それは一面に過ぎず特に質の評価に繋がらないという批判もあるが、定量・定性の両面を見る政策評価等において欠くべからざる指標であることに異論の余地はない。以降の予算取りに影響するため、ウェブサービスに関わる者は概してアクセス数を気にするものだ。

アーカイブについても同様で、アクセス数が増えないという嘆きをよく耳にする。そして、そうしたサイトを眺めると、いくつかの共通点が見えてくる。

2. 1 解説が足りない

これはアーカイブにありがちな失敗だ。アーカイブが抱えるコンテンツの数量はウェブサービスの中でも突出している。そのせいか運営者はコンテンツの網羅に執心しコンテンツが集まるとそれだけで安心してしまいがちだ。また、運営者の多くがその道の専門家であるが故に素人である一般利用者の当惑に思いが及ばないということもある。さらには、予見を持ってほしくない、余計な脚色をせずあるがままに提示したいという思い遣りのようなものもある。加えて慢性的な人手不足も影響している。

しかしながら、説明の足りないサービスには、「素人は来るな。知らないやつは見るな。」という高慢さがついて回る。利用者は当惑し悪印象を持ち帰る。それでは困る。アーカイブのサービスにも、不案内な人を迎えて面白さを伝えようとする仕掛けと努力とが必要だ。その点でアーカイブは他のウェブサービスと何ら変わるところがない。

2. 2 更新が足りない

もう一つアーカイブにありがちな問題は更新が足りないことだ。極端な例だが、開設時に用意したアプリケーションだけで5年間集客できると思い込んでいたという反省の弁を何度か耳にした。

ところが、ウェブサービスのアクセス数には、更新直後急激に立ち上がってそれ以降は漸減するという傾向がある。更新をしなければ早晩利用者は減っていくのだ。

こう話すと「そうは言ってもアーカイブへのコンテンツ追加は中堅どころですら年に数回だ」との反論を受ける。ウェブサービスの更新をコンテンツの追加と決めてかかっているわけで、ここが誤りのポイントだ。利用者からすれば、新しい気づきが得られればそれは即ち更新である。アーカイブのレコード数自体は変わらなくとも、コンテンツをピックアップして関連付ける、解説の観点を変える、といった具合に何か新しい発見を提供すればそれは十分に来訪の動機となる。

図１は定期的に解説文を掲載しているとあるアーカイブの直近３か月のアクセス推移だ。点線がアクセス数、実線が６日移動平均線である。

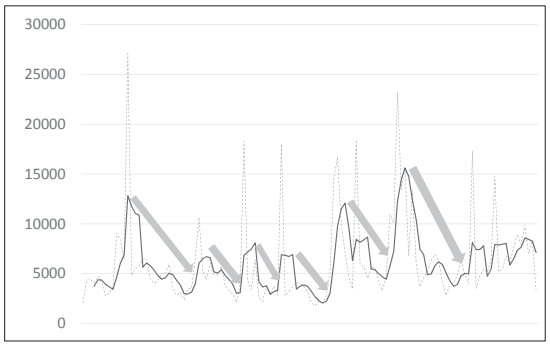

図１　アクセス数の遷移（例）

１日平均5000アクセスのアーカイブだが、解説記事の掲載日にはアクセス数が３倍～５倍に跳ね上がり、その後漸減していく様子が見てとれる。解説の掲載がサービスの成功に寄与していると言えるだろう。

ただし、それは言うに易く行うに難い。記事内容を考えるだけでも手間だ。せめて更新は楽に行いたい。日常業務として継続させるなら、ピックアップ、ギャラリー、ブログといった、簡便な更新を支える機能がシステムには必要だ。

さて、このような日常的な更新のほかに、リニューアルや新規アプリの追加といったもう少し大きな更新もある。ウェブサービスは日進月歩、時代が次々新しい機能やコンセプトを要求してくる。リニューアルを２～３年もサボったウェブサービスは古臭く見えてしまうものだ。

アプリケーションへの要求は時代とともに変化する。それをサービス導入時に予見することはできない。しかしながらその準備はできる。

そもそもアーカイブはコンテンツの供給元という役割を持っている。外部のサービスやアプリケーションにコンテンツを供給する仕組み（API）が整っていれば、追加したアプリケーションにコンテンツを扱わせることができる。しかしながらこのような一見目立たない機能を端折っているシステムも多く見受けられる。システムの導入要件に加えることをお勧めする。

2. 3 広報が足りない

公共的なサービスにおいて、広報はセンシティブな活動だ。特にSEO（Search Engine Optimization：検索エンジン最適化）についてはその実施に異論も多い。そうは言っても存在を知られなければ閑古鳥が鳴く。だから、検索サービスが利用者をナビゲートしてくれるとすれば、それは喜ばしくありがたいのだ。世間に認知されていないサービス開始当初なら尚更だ。

しかしながら広報には金も時間も努力も要る。アーカイブの運営に粉骨砕身努力する一方で広報にも気を配れと言うのはいかにも酷ではないか。IRIでは、一定の条件を備えた団体に対して、この課題についての援助を行う計画を持っている。詳細については是非問い合わせていただきたい。

3. 思った以上に費用がかかる

次に多い相談は費用についてで、大抵は、ストレージ、通信、移行、運用のいずれかのコストを圧縮したいという話だ。

3. 1 ストレージコスト

ストレージコストは、データ量・性能・地理冗長や世代管理等の付帯要件・コーデックの圧縮性能・配信の方式・クラスタギャップなど様々な条件の影響を受けて大きく変動する。従って一概に

話をするのが難しいところではあるが、それでも最低限ベンダーから他の選択肢と比較したメリット・デメリットなどの説明は受けるべきだ。これを怠ってベンダーが提案する金額を鵜呑みにした、根拠が分からないままに提案を受け入れた、といった反省を耳にする機会は少なくない。時には驚くほど高額な負担をしている例もある。

試みに、1テラバイト（A3/フルカラー/400dpi/20万枚相当）のコンテンツを扱うアーカイブをクラウド（PaaS・Blob運用・動的配信・ローカル冗長の条件）で運用するとして、ストレージコストを算出してみる。費用はベンダーの選択などにも左右されるが、国内法適用の大手クラウドベンダーで年間約3万円ほどである（執筆時点の価格）。この金額を頭の隅に留め置き、見積りとかけ離れている場合には納得がいくまで説明を求めることをお勧めしたい。

3. 2 通信コスト

静止画の配信を中心とするアーカイブを運用する場合、通信コストはそれほど問題にならない。しかし動画となるとそうはいかない。動画配信では大量のデータを送出する仕組みが必要になる。オンプレミスやデータセンタでの運用なら送出量に見合った設備を、クラウドであればContents Delivery Network（CDN）による分散配置を含めた費用負担を、それぞれ考えねばならない。

なお、IRIでは一定の条件を備えたアーカイブサービスに対し、ビデオプロキシの提供によってこの問題の改善を図る取り組みを進めている。興味があれば問い合わせてほしい。

3. 3 移行費用

システムを更新する際には移行作業が発生するが、その際に費用を請求されることがあるようだ。データベース上のメタデータを数MバイトのCSVファイルに出力するだけで100万円前後の費用が請求されたという事例も多い。

筆者にはどうにもこれが理解できない。データのエクスポートはアーカイブシステムが備えるべき標準の機能であるのだからその利用にあたって高額な手数料が発生するのは理屈に合わないからだ。

調達要件にこの機能が含まれていればこのようなことも起きない。くれぐれもご注意願いたい。

3. 4 運用コスト

死活監視・侵入／改竄検知・脆弱性対応にバックアップ等、人手の要る多様な作業が含まれる運用のコストは、金額が大きく悩みの種になりがちだ。一方で、これらはある程度まとめても人手への影響が少ないため、システムの規模を大きくすると単価が下がる傾向にある。要するにオンプレミス運用よりはデータセンタ運用のほうが安く、データセンタ運用よりはクラウド運用の方が安くつく。クラウド運用の中では、IaaS, PaaS, SaaSと順に費用負担が軽くなる。最も安いSaaSから検討を始め要件を満足しないようなら順次負担の重いものを検討していくなどの工夫をお勧めしたい。

なお、昨年来、Heartbleed、FREAK、SSL脆弱性、自己証明書攻撃などコンピュータシステムの脆弱性に起因する社会問題が相次ぎ[1)]、その影響からセキュリティパッチ適用作業のコストが膨らむ傾向にある。特にベンダーからの継続的なセキュリティパッチ提供がないオープン系のOSでは、コミュニティへの参加・パッチ入手・動作確認、パッチ適用など一連の作業を継続する負担が増大した。一方でMicrosoft AzureのPaaSのようにセキュリティパッチ適用をほぼベンダー任せにできるような環境もある。システム調達の際には初期導入コストだけではなく運用コストについても十分に検討しておきたいところだ。

4. やりなおしが多い

アーカイブ対象の資料はほとんどが貴重なものだから、何度もスキャンして負荷をかけるようなことはできるだけ避けたいものだ。たとえそうではなくても2度手間を避けるに越したことがない。

ところがコンテンツデータの作成現場では意外なほどやり直しが多く発生する。要求仕様の不備・機材の性能不足や調整不足・作業者の不注意・非可逆圧縮による劣化など原因は様々だ。

4. 1 要求仕様の不備

技術の進歩や用途を勘案せず色数や解像度などの仕様を決めると後になって作り直しの憂き目を見る。良く知られた事例に、全ての台帳を白黒二値で記録したとある大規模プロジェクトがある。色の弁別ができないと赤書きなどの書き込みが下の文字と重なって弁別不能となる。そのため結局全ての台帳をスキャンしなおす羽目に陥り膨大な追加費用を計上した。

失敗と言うには酷な例もある。ここ30年ばかり解像度は400ppiというのが電子化の半ば常識として語られてきた。オフセット印刷の資料は最低でも175線で印刷されている、これを潰れなく再現するにはその倍である350ppi以上の解像度が必要になる、というのが根拠だ。いわば入力だけを考慮して決められた仕様である。ところがディスプレイの高解像度化が進み800ppiを超えるスマホが販売されるに至って、アーカイブの解像度を見直す出版社が増えてきた。出力のほうも考慮する必要が生じたのだ。

技術は日進月歩だ。それにつれて要求水準も上がる。だからこのようなやり直しを根絶することはできない。しかしながら、保存やサービスなど電子化の工程を整理しなおすことでそのリスクを減らすことは可能だ。

できるだけ良い状態でデータを作ってこれを保存用のデータとし、そこから運用しやすい解像度や色数のデータを取り出してこれをサービス用とする。サービス用のデータは時代の要求に合わせ適宜作り直す。滅多に使われることのない保存用のデータは、低速で安価な媒体に記録しておく。

保存用のデータを作る場合、やってはいけないのが非可逆圧縮を適用することだ。非可逆圧縮は目立たない部分を捨てることによってデータ量を減らす技術だ。つまり非可逆圧縮を適用するとデータは壊れる。圧縮はデータを保存するたびに適用されるので、その都度データは壊れていく。たとえばJPEGは非可逆圧縮だから、そのデータに拡大・縮小・移動・切り出し等何らかの操作をして保存をすると、そのたびに絵は汚れていく。

このように、派生元となるデータに非可逆圧縮は不向きである。そのため保存用のデータには可逆圧縮を適用する。可逆圧縮は圧縮性能が低いためデータ量は元データの概ね半分にしかならないが、絵が壊れていく心配はない。なお、可逆圧縮と非可逆圧縮の両方に対応し、保存用・サービス用の両方にシームレスに使えるJPEG 2000という便利なコーデックがあり、各国の中央図書館ではアーカイブにこの技術を採用している[2),3)]。

これに関連して一つよく目にする失敗を紹介する。それは保存用データのフォーマットをPDFにする場合に発生する。PDFというのはコンテナ(容器)に分類されるフォーマットで、静止画についてはJPEGなどのデータをそのまま格納する仕組みとなっている。さきほど説明したとおりJPEGは非可逆圧縮なので保存用には向かない。工夫すれば可逆圧縮のJPEG 2000データを格納するなどして保存用のPDFデータを作ることも可能なのだが、現実にはこの失敗を実によく目にする。お気を付け頂きたい。

4. 2 機材の性能や調整の不足

スキャナなどの機材に問題があると、画像が潰れる、色味がおかしい等の問題が頻発してうんざりするほどのやり直しが発生する。

これを防ぐためには、電子化に先立って使用予定の機材を評価するのが良い。筆者にも立ち会う機会があり驚くような事例も経験した。RGB各256階調を謳いながら実は各60階調のセンサを使っていた明らかな粗悪品、そこまで悪質ではなくても、ダイナミックレンジが狭い、色かぶりがある、解像度が足りない、ピントがあっていない等々、実に様々な問題に行き当たる。

評価は事前に入力業者からサンプルデータの提出を受け、問題のないことを確認する形で行う。こうして問題ないことを確認された業者だけを対象に調達を行えば間違いがない。評価にはJIS X 6933 No.2等、基準データが添付された標準規格のチャートを使うのが良い。なお、作業期間中にも定期的に同様の確認を行うことをお勧めしたい。

4. 3 作業者の不注意

人為的なミスも少なくない。良心的な業者は資料を注意深く扱うが、そうではない業者もいる。資料を壊して出入り禁止になった噂を聞き確認してみるとほぼ同じ.業者だったりして、規模の大小や所在地には関係がないようだ。要するに作業者の経験と誇りの問題なのだろうと思う。

ほこりといえば、評判の悪い業者が提出するサンプルデータには埃が目立つものだ。ある案件で写っている埃の数を実際に数える機会があったが、多くとも60個程度の埃を数えた中で、ある業者だけが300個を超える記録を出していた。チャートの管理が悪いためにそんな事態を招いたのだろうが、資料の管理も推して知るべし。想像すると背中を冷たいものが走る。

4. 4 非可逆圧縮による劣化

仕様が適切で、機材のコンディションは良好、作業者も良心的。それでもやり直しは発生する。サービス用データの画質が低すぎる場合がままあるのだ。

配信の負荷を下げるためサービス用のデータには非可逆圧縮を適用する。先にも述べたように、非可逆圧縮は目立たない情報を捨ててデータ量を減らす技術だ。ここで問題になるのは、どこまでなら目立たないか、どこまでなら捨てて良いか、ということだ。情報を捨てすぎるとデータが壊れて使い物にならなくなる。

すぐに思いつくのは、安全そうな圧縮率を目安にすることだ。しかしながら、圧縮率と絵の壊れ方の関係は一定ではなく、対象ごとに大きく異なってしまうことが知られている。

一律の圧縮率を適用すると一部のデータが壊れすぎる。壊れたデータを使ってサービスを運営するわけにはいかない。しかしながら、どのデータも壊れないような低い圧縮率を適用すれば今度はデータが無駄に大きくなる。だから、圧縮率を基準にする場合には、結局全てのサービス用データを目視で確認し、画質の悪いものについては圧縮率を変えてやり直すことになる。

目視確認の最大の問題はひと工程多いことで、これがスケジュールや経費に影響する。そこで、特に大規模なプロジェクトではStructural similarity（SSIM）やPeak signal to noise ratio（PSNR）といった客観的な画質指標[4)]を導入して目視検査の工程を省略する。SSIMは原本との類似性を％で表す指標でありPSNRはノイズの割合をdBで表す指標だ。

5. オープンデータと言うけれど

オープンデータが花盛りだ。知的財産戦略本部、総務省、経済産業省といった中央省庁が旗を振り、一部の地方公共団体も環境整備に乗り出した。WWWを発明した聡明な科学者の提唱ということもあってかウェブ全体がオープンデータに肯定的だ。歓迎すべき状況であるが、黎明期ということもあってこれにまつわる失敗談や相談も少なくはない。

5. 1 Google MAPのAPIが突然使えなくなった

大手のウェブサービスはさまざまな機能をAPIとして提供している。中には自前で構築すると莫大な費用のかかる機能を無料で使わせてくれるありがたいものもある。これに着目してGoogle MAPやFacebookのAPIを利用するウェブサービスも最近になって随分増えた。Web上のいたるところで見かけると言って過言ではない。

しかしながら有難がってばかりもいられない。無償のAPIは利用者に対して責任を負う立場にないのだ。サービスを行う者がこれを忘れると大変な目に合う。

かつてGoogle MAPはAPIのバージョンを上げる際にインタフェースを大きく変更したが、その影響を受けて複数のアーカイブがサービスを継続できなくなったことがある。この場合は変更だけだったが、突然廃止されたAPIも少なくない。一方で一定の頻度を超えて利用すると料金が請求されるような利用条件のついたAPIもある。

こうしたリスクを持つAPI利用だが、それを補って余りある利便性があるために敬遠してしまうのは惜しいものだ。リスクの存在を踏まえ、必要な対策を取った上で注意深く利用していきたい。

まずは利用条件をきちんと把握することだ。利用したいAPIが費用などのリスクを持たないことを確認する。利用条件は度々更新されるので定期的にチェックする。また、似たような機能を持つ別のAPIを使ってサブのサービスを用意すれば、APIの変更や廃止があっても影響のないほうのAPIを使ってサービスを継続できる。

そして何より大切なのは情報を交換し合うコミュニティに参加することだ。IRIでも大手API提供サービスの協力を得て情報交換の場を設けている。興味があれば問い合わせてほしい。

5. 2 パーマリンクが使えない

たいていのアーカイブサービスには提供中のコンテンツを参照するためのパーマリンクのような仕組みがある。コンテンツ、表示位置、解像度、サイズなどを特定する情報をひとつのURI（Uniform Resource Locator：統一資源識別子）にまとめることで、ブログのような外部のコンテンツからアーカイブ内のコンテンツを参照できるようになる。サービスの運営者からもサービスの利用者からも喜ばれる大変便利な仕組みと言える。

ただし、これにはひとつ悩みがある。アーカイブのサービスと利用者側のサービスとで継続期間に差があることだ。利用者側サービスのほうが早く終了するのなら問題はないのだが、アーカイブ側の継続期間が短い場合にはリンク切れになってせっかくコンテンツを紹介してくれた利用者に迷惑をかけることになる。5年ごとにシステムを入れ替える公的なアーカイブでは定期的にこの問題に直面する。

これを避けるには、今のところパーマリンク移行機能を提供するシステムを利用するしか方法がない。今後は制度的な取り組みにも期待したいところだ。

5. 3 再利用条件が守られない

オープンデータの普及に合わせCreative Commons[5]の認知度が上がり、CC:BYを明記した公共サービスも増えた。これを利用し、Office系のアプリケーションが、オンライン画像をドキュメントに挿入する機能を提供したりもしているからご存知の方も多いだろう。

さて、その機能を使ってオンライン画像を二次利用する場合に、出典が記載されないことがある。CC:BYとは「二次利用していいからその代わりに出典を明示してね」という約束だ。Creative CommonsのライセンスはCC0以外全て最低限CC:BYを含んでいる。それなのに少なからず出典が見当たらない。おそらくは忘れられているのだ。

Creative Commonsのライセンスマークは大抵ウェブページの目立たないところに張られている。そのため悪意がなくともライセンスの転記を忘れてしまうことが発生し得る。そうしてライセンスとはぐれてしまったコンテンツが、さらに別の場所で二次利用される。自分のものと主張する輩が現れるかもしれない。このようにして迷子になったコンテンツを探し出して、ライセンスしなおすことは容易ではない。

この問題に対するひとつの解がISO 16684として標準化されているExtensible Metadata Platform（XMP）だ。XMPは画像や映像などのコンテンツデータに付帯的な情報を組み込む規格である。Creative Commonsでは、XMPを使ってコンテンツデータにライセンスを付与する取り組みを進めている。Creative Commonsを使った二次利用許諾を計画しているのであれば、是非XMPの導入を検討してほしい。

5. 4 連携の進め方がわからない

この頃はNDLサーチと連携したいという相談をもらう機会も増えた。この連携はOpen Archives Initiative Protocol for Metadata Harvesting（OAI-PMH）[6]という標準プロトコルを介して実施する。OAI-PMHを搭載するアーカイブのシステムは多く、これを使えば少ない負担で連携の実現が可能だ。

NDLサーチ以外の図書館システムとの連携ではOpenSearch[7]というAPIを使うことが多い。こちらも標準的なプロトコルで、搭載するアーカイブシステムも少なくない。システム選定時には是非確認してほしい。

一方で、TwitterやFacebookのようなSNSとの連携を希望する向きもある。ここではOpen Graph Protocol（OGP）というプロトコルを使う。こちらもほぼ業界標準であり、これを搭載するシステムも増えている。

6. 結びにかえて

知的資源イニシアティブは「知的サービス研究会」を母体に2003年にNPO法人として設立された団体だ。様々なMLAに関わりを持つメンバーが、それぞれの問題意識に従って活動を進めている。図書館総合展での Library of the Year の主催、ISOから委託されたJPSEC Registration Authorityの運営、文化庁から委託された調査事業の実施等、活動は多岐にわたる。

また、IRIでは、Officeアプリ、セキュリティ、マルチメディア、ビデオ、グループウェア、広告、アーカイブ、クラウドなど、様々な分野を代表するアプリケーションベンダーの協力を得てMLA関係者の活動環境整備と情報共有にも取り組んでおり、特にデジタルアーカイブへの敷居を下げるための活動を進めている。

当会の活動に興味をお持ちの方は是非コンタクトを頂きたい。

ウェブサイト　http://www.iri-net.org/
メールアドレス　info@iri-net.tokyo

謝辞

調査編集協力：株式会社寿限無CTO　大澤　浩
査読：秋田県立図書館副館長　山崎博樹

（おかもと　あきら）

参考文献

1）上原孝之. 情報セキュリティスペシャリスト. 2016年版, 翔泳社, 2015, 721p, 情報処理教科書.
2）Edward M. Corrado et al. “Digital Preservation for Libraries, Archives, and Museums”. Rowman & Littlefield, 2014, 294p.
3）Federal Agencies Digitization Guidelines Initiative. “JPEG 2000 Summit Workshop”. http://www.digitizationguidelines.gov/resources/jpeg2000.html 他多数（参照2015-11-27）
4）杉本修. “符号化画質”. 電子情報通信学会知識ベース 2 群 5 編 9 章. http://www.ieice-hbkb.org/portal/doc_02_05_log.html（参照2015-11-27）
5）Creative Commons. http://creativecommons.org/（参照2015-11-27）
6）Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0 http://www.openarchives.org/OAI/openarchivesprotocol.html（参照2015-11-27）
7）opensearch.org. http://www.opensearch.org/（参照2015-11-27）

失敗に学ぶデジタルアーカイブ　～アーカイブ運営のノウハウを共有する～

岡本　　明（NPO法人知的資源イニシアティブ 理事、株式会社寿限無 代表取締役）

NPO法人知的資源イニシアティブ（以下、IRI）は、知的情報資源の収集・蓄積・利用に携わる・関心を持つ個人・団体・機関が会合や研究会を開催し、研究を行い、それらに関する啓蒙や提言の公表を行う特定非営利法人である。本稿では、アーカイブ運営の悩みや失敗事例を、1）アクセス数、2）コスト、3）電子化役務、4）オープンデータ関連、の4つの観点から取り上げ、回避策を考察するとともにそれに向けたIRIの活動を紹介する。