ディジタルアーカイブと電子文書の保存

筑波大学・図書館情報メディア研究科・知的コミュニティ基盤研究センター　　杉本　重雄

1.　はじめに

現在、我々が作り出すほとんどの文書はワープロなどを用いて電子的に作られている。配布、流通や出版のために紙に印刷することも多いが、最近ではメールや Web を介して電子的にのみ流通していることも多い。メールや Web の文書を読むためだけに紙に印刷し、読んだ後は紙を捨て、保存は電子版でということも多くある。これはある意味では「文書の電子化とペーパーレス化の進展」を表していると思う。その一方、いつの間にか以前作った文書が読めなくなっているという問題にも遭遇する。

電子文書、特にもともと電子的に作られ、電子的に流通し、電子的な環境で利用する文書の保存が、ネットワーク情報化が進む社会にとって重要な問題であることは広く認められている。また、この問題は完全な解決が難しい問題であるということもよく知られている。解決が困難な問題である第一の理由は、文書の保存が求められる期間の長さと、電子文書を利用するために必要な機材やソフトウェアの寿命との違いである。また、電子文書の場合は、保存対象となる文書の収集の面においても紙の資料とは異なる困難さを含んでいる。

筆者は、ここ数年、国立国会図書館の納本制度審議会でのネットワーク系出版物の収集に関する委員会や、内閣府における電子媒体による公文書等の管理・移管・保存のあり方に関する研究会等、電子文書の保存に関する議論に参加する機会を得た。また、ディジタルライブラリやメタデータに関する研究活動においても、電子資料の長期保存に携わる機会を得た。本稿では、こうした活動の中から得た自分なりの理解を中心に述べたい。また、これはきちんと裏打ちのされたものというよりは筆者の思いに基づくものであることをはじめにお断りしておきたい。

2.　ディジタルアーカイブ（Digital Archive）

ディジタルアーカイブということばがよく用いられている。このことばは、文化遺産を電子化して提供するものの場合のように、現実世界の「もの」を電子化し、蓄積したものを表す場合に用いられることもあれば、いろいろな電子文書を収集し、蓄積したものを表す場合もある。文化遺産の電子化の場合、図書や manuscript に限らず、さまざまな美術館や博物館資料に加えて、遺跡そのもの、さらには芸能や工芸技術などの無形文化財の保存のための電子化も行われてきている[1]。このようにディジタル

[1] “ 文化遺産オンライン ”, http://bunka.nii.ac.jp/（2006 年 9 月アクセス）

アーカイブにはいろいろな目的のものがある。以下本稿では、原資料の電子化によって作られたものをも含め、電子文書の収集と蓄積、長期保存という観点を中心に述べる。

Web 上に公開された文書を収集し保存する Web アーカイブ、図書館等が収集した電子文書を長期に保存するためのアーカイブ、組織の中で規則にしたがって組織の文書を保存するためのアーカイブなど、電子文書のアーカイブにもいくつかの型がある。電子文書といってもワープロや表計算ソフトで作成したテキストと図表のみのものから、リンク構造を持つもの、動画や３次元画像までを含むものまでさまざまである。また、電子文書には、大別して CD 等のパッケージに入れられて配布・利用するもの(パッケージ系資料）と、サーバに蓄積してネットワーク上で利用するもの（ネットワーク系資料）がある。本稿では、パッケージそのものの保存に関しては考えず、内容（コンテンツ）の保存という観点からこれらを特に区別せずに考えることにする。

ネットワークに接続されたディジタルアーカイブによって、我々はいつでもどこからでもいろいろな文書や資料を利用できるようになった。一方、電子資料は、紙やものの資料と異なり、電子資料を利用するための機材やソフトウェアを必要とする。そのため、蓄積されたコンテンツを長期にわたって利用可能な状態にとどめるには資料そのものを保存するのみならず、それを利用するための機材やソフトウェアも含めて保存する必要がある。しかしながら、速度の速い技術の進化の中で機材やソフトウェアを含めた保存は容易ではない。このことが電子資料、特にもともと電子的に作成され電子的な環境での利用を前提とした資料（Born Digital 資料）の長期保存を困難にしている。

3. ディジタルアーカイブ構築のための要素

本節では、収集、蓄積と保存、組織化と利用の視点からディジタルアーカイブに関して考えてみたい。

3.1 収集

ディジタルアーカイブ構築のための資料収集の方法は以下のように大別することができる。

・ネットワークを介した資料収集
  ・Web などネットワーク上に公開された資料の収集
    内容に基づき資料を選択的に収集する方法と、指定した範囲の資料を網羅的に収集する方法がある。
  ・組織毎の資料収集など、資料の提供者の方針を反映した収集
・オフラインでの電子ファイルの収集（パッケージの収集)
・「もの」の電子化による収集

Web 上の資料の収集には一般に収集ロボットによる自動収集方式が用いられる。Web 上での収集には、資料の識別のために一般に URL (Uniform Resource Locator)

が用いられる。Web 上の資料は更新されるものが多いため、同一の URL からの収集を繰り返すことになる。一般には、資料の更新と収集の同期がなされない。そのため、更新毎の内容をすべて収集するといった場合には目的に合わせた仕組みを用意する必要がある。また、Web 上に公開された資料であっても収集ロボットによるアクセスを受けつけないものや、自動収集に対応しないデータベースに格納された資料の場合もある。それに対して、資料の提供者とアーカイブが協調して資料を収集する場合、あらかじめ決められた方針に基づく収集が可能である。また、アーカイブそのものの信頼性を高めるために、収集した資料を他のアーカイブに寄託すること、アーカイブを持つ組織の変化に対応することなどの問題に関しても考える必要もある。

3.2　蓄積と保存

電子文書の保存は、その内容であるディジタルデータをビットデータとして正確に蓄積、保存するだけでは、文書の保存の目的を果たせない。そのため、データを利用するために必要な情報を一緒に蓄積保存する必要がある。Open Archival Information System（OAIS）は、アーカイブの参照モデルを述べている。そこでは、元のデータだけではなく、再生に必要となる情報を元データに加え、さらに保存のための情報を加えて構成したパッケージの形で保存することになっている。

どのような種類のディジタル資料であっても完璧な保存ができるということは期待できない。再生のための情報や保存のための情報が的確に残されていたとしても、資料の再生のために必要なソフトウェアや機材がなくなってしまっていれば資料の再生はできない。また、外部の資料へのハイパーリンクを含む資料の場合、リンクを有効に保ち続けることは容易ではない。こうした問題は、ワープロ文書など身近にある電子文書に関しても容易に想像できる。そのため、資料の保存の際に、もとの資料が持つ機能をどの程度失ってもよいかをあらかじめ検討しておくことが求められる。

3.3　組織化と利用

メタデータを用意すれば長期保存が可能になるということはいえないが、長期保存のためにメタデータは必要不可欠である。先述の OAIS を基礎とした保存のためのメタデータ記述要素が提案されている。また、EAD や METS のように電子資料のアーカイブを指向したメタデータの標準規格が作られている。Web アーカイブを指向したメタデータの規格の検討も進められている。

一般に電子資料の保存のためには、保存対象資料の内容のみならず、その形態や保存に必要な環境、保存の履歴などの情報を記述する必要があるので、保存のためのメタデータの記述は複雑になりがちである。そのため、メタデータの記述にかかるコストをできるだけ小さくすることが求められる。ファイルの形式など記述を自動化することが可能な部分もあるが、対象資料の内容に関する記述などのコストを下げるにはできるだけ資料の作成段階でメタデータを付与することが望まれる。

資料の検索機能やアーカイブ内のナビゲーション機能など、保存した資料の利用性

を高めることも重要な視点である。こうした機能はアーカイブの内容や目的に依存する。たとえば、イメージデータとして電子化した資料の内容記述による検索機能、資料から取り出したキーワード一覧によるナビゲーション機能などは基本的に必要とされるものであろう。児童生徒・学生、成人、一般人、専門家といったさまざまな利用者に対してアーカイブの利用性を高める必要がある。その意味では、何らかのテーマからアーカイブの資料を紹介する Web 上での特別展示や、アーカイブの内容や利用方法の説明等も求められる。また、アーカイブの Accessibility（障害の有無に関わらずに利用できること）も忘れてはならない課題であろう。

アーカイブが発展するとともに複数のアーカイブの横断的な検索機能の重要性が増していくと思われる。アーカイブによって性質がさまざまであるため、それぞれの特色を生かしつつ広い範囲のアーカイブを検索するための機能が求められる。

こうしたアーカイブの付加価値を実現するには、アーカイブ毎の性質を反映しつつ、共通化と個別化の両面でアーカイブの組織化方法について検討する必要がある。

4. 公文書の電子的保存に関して

本節では、昨年度の内閣府における電子媒体による公文書等の管理・移管・保存のあり方に関する研究会での議論やオーストラリア国立公文書館の訪問調査などを通じて、筆者の学んだことについて述べたい。その多くは、よく考えると当たり前のことのようにも思えるが、自分自身の印象に残ったこととして述べたい。

⑴ 図書館のディジタルアーカイブと公文書館のディジタルアーカイブ

図書館や博物館、公文書館におけるディジタルアーカイブは、ネットワークを利用して、いつでも、どこでも、誰にでも、貴重な情報資源を提供するために、90 年代のインターネットの爆発の早い時期から取り組まれてきたものである。

図書館でのディジタルアーカイブというと歴史資料や貴重資料を電子化して蓄積したものがまず思い浮かぶ。歴史資料や貴重資料のディジタルアーカイブは公文書館でも同様な取り組みが行われている。こうした取り組みでは両者に大きな差はないと思う。また、両者ともに Born Digital 資料のアーカイブ作りに取り組んでいる。そのため、さまざまな共通の悩みを抱えている。その一方、図書館が出版された文書（公表された資料）の収集を基本とするのに対して、公文書館は業務の中で作られた文書を収集するため、収集に求められる精度、収集の対象範囲、文書の作成者とアーカイブとの関係などが両者の性質の差となって現れていると思える。また、性質が異なるといっても共通する部分を多く持つので、アーカイブを支える知識や技術の共有、サービス間での協調が求められることはいうまでもない。

⑵ 電子文書について－定義、同定・識別、その他

Born Digital 資料の保存の場合、保存の対象となる電子文書をどのように定義するかという問題がある。利用者や利用環境に応じて表示形式を変えることができること、データベースに収められた部品を動的に組み立ててひとつの文書を作ることがあ

ること、ハイパーリンクを持つことといった動的な性質は電子文書の重要な特色である。そのため、保存対象であるひとまとまりの文書とは何かということから定義しなおすことが求められる。たとえば、ある特定のブラウザ上に表示された形式を「文書」としてとらえるべきなのか、あるいはもともとサーバ上に格納されたものが「文書」なのかといった問題である。現状では、ワープロや表計算ソフトの文書など、従来の印刷物としてのイメージから連想できるものを「文書」としてとらえておけばよいと思われるが、将来にわたっては、保存の対象となる「電子文書」とはなにかに関する共通理解を作り上げていかねばならないと思われる。それには、いろいろな経験と進化の激しい情報技術への理解が求められる。

⑶ 公文書のライフサイクルと文書の流通基盤

公文書の保存は、文書が作り出されてから保存・廃棄されるまでのライフサイクル全体の中でとらえなければならない。文書の作成段階で、文書の流通や検索のためにある程度のメタデータを作っておくことで保存段階でのメタデータ作成コストが下がると期待できる。公文書の検索のための共通のメタデータ規則や分類のための基準を整備することは効率の良い電子文書の保存にもつながる。

現用段階で好ましい文書形式と保存段階で好ましい文書形式は必ずしも一致しない。現用文書が備える機能を削って保存に適した形式にしなければならない場合もある。そこでは、保存しなければならない文書内容に関する共通理解が求められる。そのため、文書の流通利用環境と文書のライフサイクル全体を見通した一貫した電子文書管理を行うことが望まれる。すなわち、文書を作る側と文書を保存する側の協調が不可欠であると思う。一方、文書管理を行うことが、技術の進歩に伴う新しい電子文書技術の導入に対する障害になることがあってはいけないと思う。

⑷ オーストラリア公文書館での調査からーまずははじめてみよう

昨年 10 月、公文書の電子的な保存に関する調査のためにオーストラリア国立公文書館を訪問した。この訪問から学んだことをいくつかあげたい。

現在行っているサービスでは、文書はネットワーク経由ではなくパッケージに入れて公文書館に送ること、ウィルス対策のために一定期間おいた後サーバに格納すること、サーバは十分セキュリティ対策がなされた場所、ネットワーク環境におくことなど実際面に対する考慮は興味深かった。また、内容の長期保存のために XML を基礎とした文書保存のためのツール XENA の開発を進めている。現時点では、どのような文書であっても対応するというものではないが、よく用いられる文書形式に対応すること、拡張性に配慮していること、メタデータと文書を一体化して蓄積することといった特徴を持つ。また、オープンソースによる開発をしており、他組織との協調を視野に入れている。こうした取り組みは、今できるベストのことを行うというものであろう。

すでに公文書は電子的に作られ、利用されているのであるから、それを電子的にアー

カイブしていくことはごく自然なことであること、また、電子文書の性質故に、電子文書の保存に関するさまざまな問題を容易に想像できるが、困難な問題があるから取り掛からないのではなく、解決できる問題から取り掛かってゆくという姿勢であるということを強く感じた。

また、オーストラリアでは、早くから公文書の組織化のために AGLS メタデータと分類語彙等を定め、それに基づく文書作成段階でのメタデータ付与を促進する努力をしてきている。こうした努力も公文書の効率的な流通と蓄積、そして保存に結びつくと感じている。

5. おわりに

電子資料の保存は電子図書館における重要でかつ解決の困難な問題のひとつである。たとえば、国立国会図書館ではこれまでも電子資料の長期保存に関する調査研究を行ってきている[2]。また、平成 17 年度に設けられた、内閣府における、電子媒体による公文書等の管理・移管・保存のあり方に関する研究会での議論は、公文書等の適切な管理，保存及び利用に関する懇談会の報告書の中にまとめられている[3]。

公文書館には、さまざまな省庁から文書が持ち込まれる。これまでは、基本的に「紙の資料」の保存を考えておけばよかったのに対し、電子資料の保存の場合には非常に多様な資料に対応しなければならないと思われる。こうした問題の解決のための努力は求められるが、すべてを公文書館だけで解決できるとは思えない。そのため、より広い範囲での協調が求められる。加えて、難しい問題があるから電子保存をあきらめるのではなく、「できることからまずはじめよう」の姿勢が求められると思う。

Web 上のサービスを人だけではなくコンピュータに提供するインタフェースを提供することが一般化している。また、利用者と利用環境の多様化も進んでいる。いつでも、誰でも、どこからでも、その環境に合わせて適切に情報提供することが求められる。こうした新しい環境に向けたサービスの実現には、いろいろな機関、組織との協調が不可欠であると思う。

内閣府における、電子媒体による公文書等の管理・移管・保存のあり方に関する研究会に参加する機会を得たことは、公文書のライフサイクルや文書管理業務に十分な知識を持たない筆者にとって、公文書の電子的保存を考える上でこの上もなくありがたいことであった。末筆ながら、同研究会でお世話になった皆様方に心からお礼を申し上げたい。

---

[2] 国立国会図書館．“ 電子情報の長期的な保存と利用 ” (報告書のほか http://www.ndl.go.jp/jp/aboutus/preservation.html （2006 年 9 月アクセス)

[3] 公文書等の適切な管理，保存及び利用に関する懇談会（内閣府）．“ 中間段階における集中管理及び電子媒体による管理・移管・保存に関する報告書 ”，2006，32p (http://www8.cao.go.jp/chosei/koubun/kondankai14/houkoku.pdf)