ディジタルアーカイブとその長期利用に関して

杉本重雄
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

1. はじめに

ディジタル情報技術の発展とともに多様なディジタルコンテンツが作られてきた。特に、90 年代のインターネットの爆発的な広がりとともに、ネットワークを通じて発信されるディジタルコンテンツが飛躍的に増えた。それとともに、ディジタルコンテンツを集積したディジタルアーカイブ(Digital Archive)が発展してきた。現在、ディジタルアーカイブということば自体、かなり広い意味で用いられていると感じられる。一方、ディジタルコンテンツの長期利用性が必ずしも保証されないことは広く知られており、ディジタルアーカイブを長期に渡って使い続けることはディジタルアーカイブにとっての共通の問題であるといえる。そこで、ここでは、ディジタルアーカイブについて、できるだけ広い視野から眺め、ディジタルアーカイブの長期利用に関して考えてみたい。

2. ディジタルアーカイブを俯瞰する

アーカイブということばは、将来に向けて、現在ある文書あるいは記録を集め、蓄積し、それを長期に渡って提供するサービス、あるいはそれを行う組織を意味する。ディジタルアーカイブの場合、ディジタル情報技術を用いて表現されたコンテンツを集め、蓄積して行うサービスととらえることができる。これまで「ディジタルアーカイブ」として、蓄積されてきたコンテンツには、電子化された文書、あるいは電子的に作られた（Born Digital）文書や記録などが含まれるのみならず、建造物や遺跡、美術工芸品などの物体をディジタルコンテンツ化したものも含まれる。そこで、ここではこれらを大雑把に分類してみよう。

1. 貴重書や歴史的文書などを提供するもの
2. 公文書（行政文書）を保存、提供するもの
3. 学術論文を収集し、提供するもの
4. Web 上の資料を収集し提供するもの
5. さまざまな文化遺産を電子化し、提供するもの
6. 電子的に出版された図書や雑誌を提供するもの
7. そのほか、教育情報資源や特定の分野の情報資源などを収集、蓄積、提供するもの

ディジタル情報資源は CD や DVD などのパッケージ型のメディアに作りこまれて提供されるものと、ネットワーク上で提供されるものがある。また、一般に、ディジタル情報資源は物理的

な資料を、スキャナなどを用いてディジタル化したものと、ワープロなどで作られた「生まれながらにディジタル形式の資源（Born Digital）」がある。前者はディジタル化による付加価値を目的にディジタル化されるものといえる。現在の我々の日常的な活動の中でワープロを使わずに文書を作ることが少なくなってきたことに代表されるように、電子的に文書を作ることは日常の一部になっている。電子的な文書の場合、リンクや画像や音声の埋め込みなど、紙の文書では実現できない機能が多く取り入れられている。また、物理的な物体から仮想現実感機能を用いて作り出された資源の場合、これらの両方の性質を持つといって良いかもしれない。いずれの場合も、情報資源の利用には適切なハードウェアとソフトウェアが要求される。このことが、ディジタル情報資源の保存を難しい問題にしている。

3. ディジタルアーカイブと長期保存、長期利用

ディジタルアーカイブを長期に渡って利用すること、あるいはディジタルアーカイブに蓄積された資源を長期に渡って保存することの重要性は疑えない。ここでは、ディジタルアーカイブとは何か、保存に関わる基本的な問題はなにかといった、基本的な項目について考えてみたい。

(1) 「アーカイブ」ということば

アーカイブは、本来、収集した資料を長期に渡って保存することを前提としてきたと思われる。しかしながら、2 節に示した 1 ～ 7 に示すディジタルアーカイブということばの使われ方をみると、ディジタル形式で蓄積・利用できるようにすることに必ずしも「長期の保存」を含意していない。しかしながら、実際にはそれなりのコストをかけて作成したアーカイブが短期間（たとえば、5 年あるいは 10 年）で利用できなくなるのはあまりに無駄が大きい。その意味では、ディジタルアーカイブには、長期に渡ってサービスを提供し続けることが暗黙の内に求められていると考えたい。

(2) アーカイブと保存

収集して、保存ならびに提供するということはアーカイブにとっての基本機能であろう。一方、ディジタルアーカイブの場合、これらを分離してとらえることができる。たとえば、図書館が持つ貴重資料をディジタル化し、ディジタル化した資料の長期保存を任務とする別の組織に送って保存を依頼する一方、ディジタル化資料を提供に適した形式に変えてネットワーク上で提供することができる。ディジタル情報資源のコピーと処理のしやすさを考えると提供の際のさまざまな付加価値サービスが可能である。

(3) 何を保存するか？

いうまでもないが、ディジタル情報資源は、紙やものと異なり、物理的に形を持つものではない。パッケージ型の資源であっても、パッケージ（たとえば、1 枚の CD）を残すこととその中身を残すこととは異なる。また、ディジタル情報資源は利用する環境や利用者に応じて表示形式を変えることができるように設計されていることもある。一方、保存に際してディジタル資料をも

とのまま残すことは必ずしも用意ではない。こうしたことは、ディジタル情報資源の保存には、ディジタル情報資源の「何を」保存すべきかを決めることも含まれると理解できる。たとえば、みばえや使い勝手（Look and Feel）も含めた完全な保存、データとしての完全な保存、テキストだけの保存、スナップショットの保存などいろいろなケースが考え得る。

(4) 保存のための文書機能、品質の低下はどの程度許されるか？
ディジタル資料、特に Born Digital 資料の保存の際に文書が持つ機能の停止や品質の低下を避けられないことがある。たとえば、ハイパーリンクを持つ文書の保存の場合、リンクが行き続けることを保証することは困難である。CAD システムで作られ、利用されている文書（たとえば、設計図）を、もとのまま保存するには CAD システムそのものの保存が求められることになる。こうした場合、保存のために機能の低下を避けることはできない。逆に、機能の低下を許すことで保存が可能になるとも言える。そこで求められるものは許し得る機能や品質低下の範囲を決めるガイドラインであろう。

(5) アーカイブは長命か？
最近の省庁の再編、自治体の統廃合や大学の統合など、従来はとても安定していると思っていた組織が簡単に変化してしまうことを経験してきた。こうした組織の改変はディジタルアーカイブにも影響する。アーカイブの高信頼化のために、蓄積した資源をアーカイブ間で移動したり、共有したりする必要がある場合もあろう。また、法律や規則の変更によりアーカイブの運用方法が変化することもある。そうした変化に対応しながらアーカイブを運営できないと、アーカイブはブラックホールのようになってしまう可能性もある。

(6) 非ディジタル資料のディジタル化とディジタル資料の保存
非ディジタル資料をディジタル化することは、原資料の保存とアクセス性の高いディジタル化資料の提供という面から進められてきている。ディジタル化にかかるコストとアクセス性を高めることによって得られるメリットに関する議論は必要ではあろうが、ディジタル化によって我々がネットワーク越しに利用できる情報資源が豊かになることは間違いない。また、提供されるディジタル化資料を基礎として作り出されるさまざまな情報や知識によって、我々の情報環境はより豊かになっていく。その一方、ディジタル化した資料であっても、適切にメンテナンスをしていかなければ利用できなくなっていく。ディジタル化資料そのものだけが失われると、その上に作り上げられたいろいろな情報資源も影響を受けてしまう。

4. Web 上の情報資源の保存　-　Web アーカイブ

Web は我々の情報基盤となっている。Web を介してさまざまな情報が発信されている。Web 上でしか発信されない情報も多くある。Web 上の情報はよく変化する。安定しない情報資源なので保存する価値がないという議論もあるが、逆にその時代を表す情報がとてもよく現れるから保存すべきであるとも言える。いずれにしても Web 上の情報資源の完全な保存は困難である。加え

て、Born Digital 情報資源に共通の保存の困難さ、違法コンテンツやウィルスなどの問題もある。

Web アーカイブは大別すると人手による選択的収集によるものと、ソフトウェアロボットを用いて行う網羅的収集によるものがある。前者の場合、内容の評価に基づいて収集されるが大規模化には困難を伴う。後者の場合、大規模化は可能であるが、基本的に内容に関わらず収集することになり、内容による評価を効率的に行うことが求められる。後者の代表とてして良く知られているものにインターネットアーカイブ（http://www.archive.org/）がある。これはインターネットを介して公開された Web ページを収集し、蓄積保存するものである。納本図書館である国立図書館の場合には、自国の、あるいは自国に関する Web ページを収集することになる。欧米の国立図書館を中心に Web アーカイブのためのコンソーシアムを作っている[1]。また、Web ページは従来の出版物とは性質が異なるため、各国で法律を作るなどして対応している。我が国の場合、国立国会図書館の納本制度審議会がネットワーク系出版物に関する答申を出しているが、現時点で法律改正にまではいたっていない[2]。

インターネット上の Web ページを網羅的に収集するシステムは、リンクが与えられないと Web ページを収集することができない。また、イントラネット内部のページのように外部からは見えないページも多くある。基本的に、こうしたページはインターネット上での Web アーカイブの対象外と考えることもできる。一方、そうしたページを発信している組織が組織自身の Web アーカイブを行う場合、いわば組織内 Web アーカイブを考えることもできる。インターネットだけではなく、イントラネットを利用した Web アーカイブも同様に考えることはできる。一方、組織のアーカイブポリシーに基づいて行う Web アーカイブも必要である。

5. 機関リポジトリとアーカイブ

現在、我が国のみならず欧米各国でも機関リポジトリ（Institutional Repository）の開発が盛んに進められている。機関リポジトリは学術雑誌の電子ジャーナル化の発展と学術雑誌の値段の高騰によって後押しされている。機関リポジトリでは論文や報告書などのコンテンツの収集蓄積と提供に目が行きがちであるが、保存の視点も忘れてはならない。

機関リポジトリは、一般に大学図書館のように安定した組織が提供している。一方、現在では、論文や紀要などの発信は誰であってもできる。また、一般の利用者にとっては Google あるいは Google Scholar のようなサービスの方が、特定のリポジトリを探し出して使うよりはなじみやすい。こうした環境を考えると、収集した資源の長期の保存は、安定した組織が提供する機関リポジトリの重要な要件であるように感じる。学習コンテンツについても同様である。

大学や研究機関では、論文にはのらないけれども面白い内容のページが提供される。こうしたページは一般には消えていく。（運よく Web アーカイブの対象になるものもあるが。）機関リポジトリを組織として行うアーカイブとしてとらえれば、組織内の Web アーカイブと機関リポジトリの組み合わせで、より内容の豊富なアーカイブになることが期待できる。

6. 公文書のアーカイブ

公文書は歴史資料として将来に残さねばならない。紙の資料はこれまで同様の方法で残していくことができる。しかしながら、電子政府化や電子自治体化の進展にともない、Born Digital の資料が増えることは疑えない。そこで、ディジタル情報資源の保存に共通の問題以外に、直感的に思いつく問題点をいくつか挙げてみたい。

(1) 電子公文書とは何かに関する定義： 電子メールや Web ページ、掲示板への書き込み、ブログは保存すべき公文書か、といった問題以外に、データとテンプレートあるいは適切なスクリプトがあれば、印刷形式や表示形式を動的に作り出すことは簡単であり、その際、何が記録として残すべきものか（もとのデータなのか、作り出された印刷形式なのか）がはっきりしない。

(2) 保存のために許される機能と品質低下の範囲： ディジタル文書の保存には機能低下が避けられない。どのような機能や品質の低下が許容範囲であるのかについてのガイドラインなしの保存は困難であろう。

(3) 原本性の保証： ディジタル資料は簡単にコピーができる。信頼できるアーカイブ（Trusted Archive）から外部に提供されたコピーについて、そのものだけで原本性を保証できるようにする必要があるのかどうか、またあるとした場合に長期間有効な保証方法はどのようなものが適切なのかといった問題がある。

以上のような問題のほかに、資料のメタデータをいかに効率よく作り出すかと言った問題もある。資料の内容に関わるメタデータに関しては、保存の現場で作成するより、資料の作成現場で作成するほうが効率的であることは疑えない。公文書の保存は公文書のライフサイクル全体から考えなければならない問題である。

7. ディジタルアーカイブとメタデータ

ディジタルアーカイブにおいてメタデータは重要な要素である。ディジタルアーカイブでは、資料の検索や内容に関する記述を与える記述メタデータ（Descriptive Metadata）、資料のアクセスや保存管理のための管理メタデータ（Administrative Metadata）、資料の（物理的、論理的）構造を表すための構造メタデータ（Structural Metadata）といった、異なる側面からの記述が求められる。

ディジタル資料のアーカイブと長期保存のためのメタデータとしては Open Archival Information System (OAIS)参照モデル[3]を基礎として提案されたもの、PREMIS ワーキンググループ[4]によって提案されたものがある。また、EAD (Encoded Archival Description)[5]や METS (Metadata Encoding and Transmission Standard)[6]といったメタデータスキーマも提案されている。実際には、アーカイブする資料の種類、アーカイブの目的などによって適切なスキーマが選択されることになる。また、一般に、保存のためのメタデータ記述は項目が多くなる。管理情報や構造情報など付与の自動化が行いやすい部分もあるが、メタデータ付与の効率化が求められる。

8. おわりに

ここでは、ごく簡単にディジタルアーカイブの長期利用に関して概観してみた。ネットワーク情報化社会は今後もますます進んでいくことは疑えない。電子出版物もますます増えていくことは疑えず、行政情報やサービスの電子化も進んでいく。我々の周りにディジタル情報資源がどんどん増えていくことは間違いないにもかかわらず、それらを適切に将来に残していくことに関しては心もとない状況であると思う。筆者は数年前にディジタルアーカイブについて解説記事を欠いたことがある[7]。そのころと比べると理解は進んできたように思う。その一方、90 年代からディジタル情報資源の保存の重要性と難しさは言われてきたが、それが現実の問題として現れてきたという状況のように思う。

参考資料

[1] International Internet Preservation Consortium, http://netpreserve.org/

[2] 国立国会図書館納本制度審議会，ネットワーク系電子出版物の収集に関する制度の在り方について，2004，http://www.ndl.go.jp/jp/aboutus/data/a_toushin_2.pdf

[3] Reference Model for an Open Archival Information System (OAIS), http://public.ccsds.org/publications/archive/650x0b1.pdf

[4] PREMIS Final Report, http://www.oclc.org/research/projects/pmwg/premis-report.pdf

[5] Encoded Archival Description, http://www.loc.gov/ead/

[6] Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/

[7] 杉本重雄, Maria Luisa Calanag. デジタルアーカイブとメタデータ, 人工知能学会誌, Vol.18, No.3，pp.217-223，2003