アルケイア―記録・情報・歴史―
第2号　2008年3月　*1-17*頁
南山大学史料室

デジタルアーカイブ構築の技術的課題について
―ボーンデジタルコンテンツとWebアーカイブ―

河 野 浩 之

Technical Problems in the Development of Digital Archives: Born-Digital Contents and Web Archives

KAWANO Hiroyuki

archeia: documents, information and history
No.2　March, 2008　pp.*1-17*
Nanzan University Archives

デジタルアーカイブ構築の技術的課題について

―ボーンデジタルコンテンツとWebアーカイブ―

河 野 浩 之

“Article 3 - The threat of loss: The world's digital heritage is at risk of being lost to posterity. Contributing factors include the rapid obsolescence of the hardware and software which brings it to life, uncertainties about resources, responsibility and methods for maintenance and preservation, and the lack of supportive legislation. Attitudinal change has fallen behind technological change.”

（UNESCO デジタル遺産の保護に関する憲章[1]より抜粋）

1章　はじめに

近年の電子情報技術の発展は著しく，記録媒体の密度向上は凄まじい．記録媒体と再生機器の市場環境の変化も劇的であり，多様な媒体へと記録された内容の再現可能性はごく短期間に失われる危機に見舞われている．しかし，これらの多様な記録媒体に記録された，記録されつつある，記録されるであろう情報を，我々は，人類の知の営みの成果としてアーカイブ（保存，収集，格納，組織化，提供）しなければならない．加えて，過去のアーカイブにある「知の成果」をも，新たな記録媒体に格納し直すことにより，情報通信技術を利用し，その共有範囲を拡大し，伝達時間の短縮が可能になる．情報化社会へと急速に変貌を遂げつつある現在，「知の成果」をデジタル化しネット

ワーク共有することは社会的な急務となっている．

例えば，欧州デジタル図書館のプロトタイプサイトである“Europeana”(http://www.europeana.eu)は2008年11月に始動する予定で，「ヨーロッパの図書館，文書館，美術館などから集められた，少なくとも200万点のデジタル化された資料（本，写真，地図，音楽，映画など）へアクセス可能になる」[2]．また，世界各国において，コンピュータネットワークを駆使し，既存のコンテンツをデジタル化するのみならず，ボーンデジタルコンテンツの流通を促進するプロジェクトが目白押しである．

以下，本稿は，デジタルコンテンツのアーカイブサービスの動向と，その技術的課題を中心とした紹介を行う．2章では，ボーンデジタルコンテンツを扱うWebアーカイブの現状を中心に，世界各国のデジタルコンテンツのアーカイブサービスの動向を紹介する．3章では，デジタルアーカイブの技術的課題を述べる．特に，メタデータフォーマットとデータ交換プロトコルに関する課題は，今後，複数機関のアーカイブを相互連携する上で鍵を握る技術である．また，アーカイブシステム構築に必要なソフトウェアについても併せて触れる．そして，4章を結びとする．

2章　デジタルアーカイブとWebアーカイブ

本章では，デジタルアーカイブについて，コンテンツの量と特性，各国機関プロジェクト，日本の動向について紹介する．

2．1　アーカイブ対象となるコンテンツ

アーカイブ対象として代表的な書籍を，スキャナーでデジタル化してアーカイブした場合，およそどの程度の記憶容量が必要になるかを概算した結果を図1に示す．

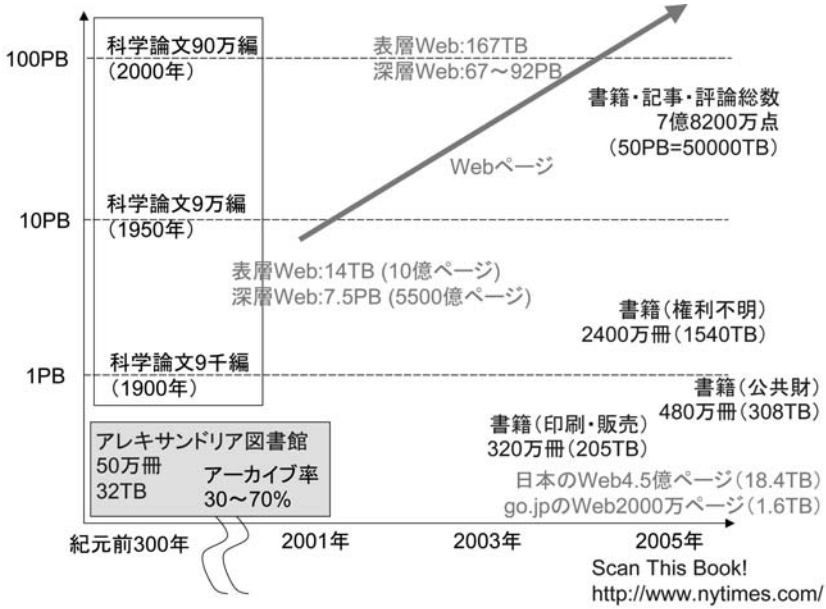

図１　各種アーカイブ対象のデジタルデータサイズ換算

“Scan This Book!”[3]によると，世紀前のアレキサンドリア図書館では，同時代に存在した書物の約30%から70%程度に相当する約50万冊をアーカイブしていたと推定しているので，データ量を32TB(1TB=1000GB)と概算する．加えて，現在，印刷･販売され流通する書籍，著作権が切れて公共財となった書籍，権利関係が不明な書籍，過去に出版された書籍・記事・評論などを，デジタル化した場合のデータ量も記入した．科学論文などに増加傾向があるものの，印刷物のデータ量はWebに直接掲載されるボーンデジタルコンテンツのデータ量に比較すると格段に少ない．実際，2001年のWebサイトのデータ量が，ファイルとして格納された表層Webが10億ページあり，データ量が14TB(1TB=1000GB)，データベース等により提供される深層Webが5500億ページあり，データ量が7.5PB(1PB=1000TB)と莫大である[4]．また，市場で販

売されるデジタル機器に蓄積できるデータ量は，2006年の161EB(1EB=1000PB)から，2010年には988EBに成長するとの報告[5]より，今後もボーンデジタルコンテンツが同等以上のペースで増大すると予測される．

加えて，デジタルコンテンツの更新・削除が頻繁に行われるため，「Webページの平均寿命は75日」(B. Kahle, “Archiving the Internet,” 1997),「学術論文引用URLは4年で半減」 (D. Spinellis, “The Decay and Failures of Web References,” 2003)などと論じられており，ボーンデジタルコンテンツの重要性が高まる中，消え行くのみという危機的な状況に陥りつつある．

しかしながら，これらのボーンデジタルコンテンツに関する興味深い議論が，現在使用されている言語分布の調査，作成されながらも見つけることのできないコンテンツの存在など様々な文脈からなされている．例えば，ネットワーク上のWebページを記述する言語にする調査報告[6]や，検索可能なWebページ数と検索可能な言語の調査報告[7]などが参考になる．よって，これらの議論の根本である資料を失わないためにも，網羅的で無くともボーンデジタルコンテンツのアーカイブは重要である．

２．２　世界のWebアーカイブの動向

本節では，世界各国で行われている幾つかのWebアーカイブ関連プロジェクトを，文献[8]を参考に表１に示す．その一部については，デジタルアーカイブへの取り組みも併せて紹介する．

現在，デジタルコンテンツを長期保存する技術的調査や研究，法制度や組織整備が，世界各国で活発に進められている．中でも，インターネット上のWebサイトのアーカイブの取り組みにおいて，Internet Archiveの果たす役割は大きく，各国の主要図書館を中心とする国際的組織IIPC (International Internet Preservation Consortium，http://www.netpreserve.org/)の主導的役割を担う．また，米国議会図書館は，American Memoryコレクションや，大統領選挙や911関連サイトをアーカイブするMinervaを提供している．なお，米国の著作権にフェアユース規定があることから，多数のアーカイブが運用されて

いる.

表１：各国のWebアーカイブとデジタルアーカイブ

国	機　関	プロジェクト名	開始年
米国	議会図書館	American Memory	1996
	http://memory.loc.gov/ammem/index.html		
米国	議会図書館	Minerva	2000
	http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html		
米国	Internet Archive	Internet Archive	1996
	http://www.archive.org/index.php		
スウェーデン	スウェーデン国立図書館	Kulturarw3	1996
	http://www.kb.se/soka/internet/sv-webbsidor/om/		
デンマーク		NetArchive.dk	2001
	http://netarchive.dk/index-en.php		
オランダ等	Networked European Deposit Library	NEDLIB	2000
	http://nedlib.kb.nl/		
オランダ	オランダ王立図書館	e-depot	2000
	http://www.kb.nl/dnp/e-depot/e-depot-en.html		
イギリス	英国図書館	Domain.uk Collect Britain	2001
	http://www.collectbritain.co.uk/		
フランス	フランス国立図書館	Gallica	2001
	http://gallica.bnf.fr/		

オーストラリア	オーストリア国立図書館	PANDORA	1996
	http://pandora.nla.gov.au/		
オーストリア		AOLA	2001
	http://www.ifs.tuwien.ac.at/~aola/		
ドイツ	ドイツ国立図書館	KOPAL	2003
	http://kopal.langzeitarchivierung.de/index.php.en		
カナダ	カナダ国立図書館・公文書館	Library and Archives Canada	
	http://www.collectionscanada.gc.ca/index.html		
中国	中国国家図書館	国家デジタル図書館プロジェクト	2002
	http://www.nlc.gov.cn/en/indexen.htm		
台湾	国家図書館(台湾)等	NDAP, Taiwan	2002
	http://www.ndap.org.tw/index_en.php		
韓国	韓国国立図書館	韓国国立デジタル図書館	2008 予定
	http://www.nl.go.kr/nlmulti/index.php?lang_mode=j		
日本	国立国会図書館	WARP	2002
	http://warp.ndl.go.jp/		

北欧各国はデジタルアーカイブを積極的に推進しており，例えば，スウェーデン国立図書館（Kungliga Biblioteket: KB）では，1661年に開始した納本制度を，電子出版物収集のために1994年に改正し，電子出版物の収集を実施している．

オランダは，NEDLIBの中心的存在であり，オランダ王立図書館（KB）とIBMが，OAISに準拠した電子情報保存システムDIAS(Digital Information and Archiving System)を開発している．また，電子出版物のデジタルアーカイブ

も進みつつある．

ドイツでは，DIASを拡張し，ドイツ国立図書館等によるデジタル情報長期保存協同プロジェクト “kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen)” が進行している．

カナダでは，カナダ国立図書館・文書館（LAC, Library and Archives Canada）の組織再編が進められた．現在，約1億件以上，ファイル容量4TB以上のWebアーカイブをLAC館内で公開している．注目すべき点は，そのシステムの基本ソフトウェアが， 3章で触れるHeritrix (v1.12.1), nuthchwax (v0.10.0), Wayback (v0.8.0)等のWebアーカイブ構築の基本ソフトウェアを利用していることである．

また，欧米を中心とするアーカイブ構築と異なる課題であるCJK(Chinese-Japanese-Korean)言語処理を必要とする各国においても，デジタルアーカイブ構築が進められつつある．中国では，中国国家図書館（NLC）が中国国家デジタル図書館プロジェクトを推進している．台湾では，国家図書館，博物館群，大学などの参加する包括的デジタルアーカイブプロジェクト「數位典藏國家型科技計畫(NDAP: National Digital Archives Program)」が2002年から開始している．シンガポールでも，sgドメインのWebサイト収集，電子出版物やオンライン出版物の納本が進んでいる．

韓国では，2005年6月から，WebサイトやWeb文書保存も実施しており，2008年に国立デジタル図書館（NDL）を設立し，公共図書館との協力ネットワーク網の拠点構築が進む予定である．また，米国議会図書館所蔵の韓国古書のデジタル化事業などの検討も進んでいる．

２．３　日本のデジタルアーカイブ事業

国内でも，国立国会図書館以外に，国立公文書館(http://www.digital.archives.go.jp/)における公文書，古書・古文書や重要文化財等の展示閲覧，国立美術館の所蔵作品総合目録検索ならびにギャラリー(http://search.artmuseums.go.jp/, http://search.artmuseums.go.jp/yuuhokan/)，その他，国立博物館の 所蔵品検索

(http://www.tnm.jp/jp/gallery/, http://www.kyuhaku.jp/collection/collection_top.html, http://www.narahaku.go.jp/meihin/) などコンテンツのデジタル化とデータベース化が進行している．

以下は，国立国会図書館を中心に述べる．まず，CD-Rなどのパッケージ系電子出版物の長期利用保証，（著作権保護期間が満了もしくは著作権者の許諾を得た書籍及び文化庁長官の裁定を受けた書籍を中心に）明治・大正期刊行図書のマイクロフィルムをデジタル化し，NDL-OPACとの連携閲覧を可能とする近代デジタルライブラリー (http://kindai.ndl.go.jp/)，Webアーカイブを行う国立国会図書館インターネット情報選択的蓄積事業(WARP, Web ARchiving Project, http://warp.ndl.go.jp/) 等，関西館電子図書館課を中心にデジタルアーカイブが進んでいる．

特に，WARPは，インターネット上で無料公開されている電子雑誌に加えて，「立法，行政（中央省庁），司法各機関，都道府県，政令指定都市，法人・機構，大学，国際的・文化的なイベント，市町村合併に関する協議会と関連市町村等のWebサイト」を収集している．

また，将来的には，館内に収蔵している書籍はもとより，録音・映像資料等を統合するデジタルアーカイブ構築を目標にすえ，従来の書庫の役割をデジタルデータに対して行う「ストレージ層（電子書庫）」，長期保存データの管理をOAIS準拠で行う「保存システム層」，その他，収集・組織化・提供を担う各種「アプリケーション層」の実装が進む予定である．

3章　デジタルアーカイブの技術的課題

本章では，文献[9),10),11),12)]などを参考に，デジタルアーカイブの技術的課題を「保存，収集，格納，組織化，提供」の側面から簡単に紹介する．アーカイブ事業の実施においては，情報通信技術の研究開発が急速に進展する反面，著作権を含めた制度整備への「行動」の歩みが遅いこと[13)]が，喫緊の課題となっていることは，冒頭のユネスコ憲章にある通りである．

３．１　アーカイブシステムの保存モデル

現在，世界各国の多くのデジタルアーカイブシステム構築の保存モデルは国際標準規格(ISO 14721:2002)として2002年に承認された「開放型アーカイブ情報システムのための参照モデル」(OAIS; Reference Model for an Open Archival Information System)を考慮したものである．OAISは情報保存に関する総合的モデルであり，「保存計画，受入，データ管理，アーカイブ保存領域，アクセス等」[9]の概念を整理したものである．なお，OAISの機能説明において，SIP, AIP, DIPなどの用語が頻繁に登場するが，これらの用語は，コンテンツと保存情報からなる情報パッケージ(Information Package, IP)の処理過程と関係している．すなわち，システムへの投入(Submission)におけるSIP (Submission IP)，アーカイブして保管するAIP (Archival IP)，配布に用いるDIP(Dissemination IP)である．本稿では，紙幅の関係で詳細な説明は省略するとして，OAISモデルの説明は文献[9]に譲る．

また，その他に多くの先行研究があるが，ここでは， DSpace，OCLC Digital Archive，LOCKSSなどの名称を挙げておく．

３．２　デジタルコンテンツの収集

アーカイブ対象となるデジタルコンテンツをネットワーク経由で収集するには，どの程度のデータ量が存在するのか，また,どの程度の性能（リンク抽出，収集速度，データ更新判定等）が必要な「ロボット」プログラム(web robot, crawler)を必要とするか等が重要である．そこで，国立国会図書館では，アーカイブ対象となる日本のWebデータの調査を，2004年10月から2005年3月にかけて実施し，jpドメイン及びJPNIC管轄下のIPアドレスをもつWebサーバからデータを収集した．報告では，日本のWebデータ総量を約18.4TB，ファイル総数が約4億5000万ファイルと推定している[10]．

調査において，robots.txt及びロボット排除用METAタグに従った動作，JavaScriptによる動的リンク抽出精度，悪意サイトへの攻撃スクリプトを含むリンク等の通常のロボットプログラムに求められる性能改善に加えて，網

羅的アーカイブを実施する場合の問題点が明らかになった．例えば，記述内容の正確な更新判定，無限に自動生成されるカレンダーリンクや記述内容を書き換えるリンクの判定能力等である．

なお，要求されるロボットプログラムの一部機能は，wgetやHTTrack等のソフトウェアで実現されている．しかし，アーカイブシステム構築においては，IIPCで開発が進むHeritrixが鍵を握る(http://archive-crawler.sourceforge.net/)．また，Heritrixが利用するファイル格納形式であるW/ARC形式が，Webアーカイブのオープンな格納フォーマットとして検討と拡張が進んでいる．

３．３　デジタルコンテンツの格納・保存技術

デジタルコンテンツの情報が劣化しないことのみが注目されがちであるが，データを格納し再生するハードウェアやソフトウェアの変化が激しいことに対する注意が必要である．実際，パッケージ系資料の「200点のサンプル中，最新環境（WindowsXP，MacOSX）での再生に約7 割（138 点）に支障がある」との報告[9]がある．これは，フロッピーディスクやCD-Rに記録された情報を再現する場合に，生産中止となったハードウェアがある場合に読み出せない，再現環境の基盤となるオペレーティングシステムの相違やバージョン変更，ブラウザなどの再生ソフトウェアのバージョン変更などにより，再生画面の状態が異なるなどの問題を含む．加えて，ハードウェアの寿命は数年程度で，媒体劣化のため数十年程度以内に正常に読み取れないなど，デジタルコンテンツの長期保存を実施する上で重大な問題が数多くある．

そこで，媒体変換やフォーマット変換を伴うマイグレーション，異なるハードウェア上でソフトウェアを動作させて再現するエミュレーションなどの技術を必要とする．しかしながら，この種の作業やソフトウェアにも，著作権や特許権など知的所有権に関連する問題が強く関係している．

加えて，デジタルデータの改竄が容易であることから，文書格納に際して紙媒体と同様に安全に利用するための原本性保証技術も必要とする．さらに，長期保存を実現する格納媒体を必要とするが，紙媒体と同様の長期保存を実

現するには技術的な問題が多い．実際，高密度に記録されたCD-RやDVD-Rなどの媒体は，適切な温度・湿度の管理下においても30年程度の安定した保存も困難との報告が散見される[14]．

３．４　組織化とメタデータ

保存管理に関わるメタデータには，DC（Dublin Core），MARCとXMLとの相互交換の標準として開発されたMARCXML，XMLメタデータスキーマであるMETS，MARCに準拠するXMLメタデータMODSなどがある．なお，国立国会図書館のデジタルアーカイブの管理系メタデータは，PREMIS（PREservation Metadata: Implementation Strategies）に基づいたスキーマの検討を進めており，日本語の「読み」についてはMODSに設けられたscript属性を用いた記述を提案している．

また，デジタル情報保存に必要なメタデータを規定する先駆けが，CEDARS(CURL Exemplars in Digital Archives)プロジェクトである．報告において，「ファイル形式やサイズ，表示に要するソフトウェアのバージョンやパラメータ設定，オペレーティングシステムのバージョン，ハードウェアのCPU性能，メモリ容量，記憶装置，周辺機器など」の記述，その他，「識別子，受入過程や権利関係，バージョン間の関係など」を必要としたとあり，これらの属性を含むメタデータスキーマの設計の検討が進む．

もっとも，同一組織内においても，デジタルコンテンツ保存に適したメタデータスキーマの設計を模索している現状であり，表２に示すようにOPACとWARPにおいてもメタデータに相違が見られる．

３．５　アーカイブデータの提供

アーカイブシステムとしては，WaybackMachineを開発するInternet Archiveが中心となって立ち上げたIIPCを中心にWeraやnutchwax ("Nutch + Web Archive eXtensions")，(http://archive-access.sourceforge.net/projects/wayback/, http://archive-access.sourceforge.net/projects/nutch/)の実装が進んでいる．また，

データ収集から公開までの一貫した操作を支援するソフトウェアの開発も進行しており，“The Web Curator Tool Project” (http://webcurator.sourceforge.net/) がある．

その他，アーカイブデータの提供にあたっては，欧州規模の電子学位論文ポータルを構築する実験 “European e-Theses portal” におけるOAI-PMH（Open Archival Initiative Protocol for Metadata Harvesting）プロトコルを用いたハーベストが興味深い．

表２：OPACとWARPのメタデータの相違

OPAC	WARP
タイトル	書誌タイトル
著者・編者	出版者・公開者
	編者
出版地	起点 URL
出版年	収集日付範囲
件名	
分類記号 （NDC, NDLC, LCC, DDC,UDC, GPO）	NDC
標準番号 (ISBN, ISSN, CODEN, UTM, ISRN, ISMN 他）	ISSN+ISBN
書誌番号 （全国書誌番号，USMARC，UKMARC，OCLC 等）	
請求記号	メタデータ ID
各種コード(本文の言語，原文の言語，官庁，大学等)	
和図書，洋図書，電子資料，音楽録音・映像，蘆原コレクション等	コレクション種別
	NDL 資源タイプ

4章　むすび

本稿では，デジタルコンテンツに関わるアーカイブサービスの動向と，その技術的課題を簡単に紹介した．世界各国において，次世代の知識基盤となるデジタルアーカイブサービスが開始されており，デジタルコンテンツの利活用を促進する制度設計や組織整備が進んでいる．

現在，情報通信技術が短期間に大きく変化しているが，アーカイブ構築においては，コンテンツ間の関係を長期間にわたって築き上げ，コンテンツに深みを与えるメタデータに関わる課題が重要であり，今後も継続的に検討を要するであろう．特に，近い将来，図書館･文書館･博物館･美術館などに保存されている「知の成果」を相互連携するアーカイブシステムを構築する上で必須の課題となるからである．

注

1) UNESCO, “Charter on the Preservation of Digital Heritage,” 2003.
http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html
UNESCO, “デジタル遺産の保護に関する憲章（仮訳）,” 2003.
http://www.mext.go.jp/unesco/009/005/003.pdf
2) 国立国会図書館,「欧州デジタル図書館のプロトタイプサイトが“Europeana”として来秋始動」, カレントアウェアネス-R, 2008.
http://current.ndl.go.jp/node/7310
3) Kevin Kelly, “Scan This Book!,” The New York Times, 2006.
http://www.nytimes.com/2006/05/14/magazine/14publishing.html
4) “The 'Deep' Web: Surfacing Hidden Value,” 2001.
http://www.brightplanet.com/technology/deepweb.asp
5) Clint Boulton, “Billions and Billions of Gigabytes Served,” internetnews.com, 2007.
http://www.internetnews.com/storage/article.php/3663641
6) Martin Ebbertz, “Das Internet spricht Englisch und neuerdings auch Deutsch,” 2002.
http://www.netz-tipp.de/sprachen.html
7) Antonio Gulli and Alessio Signorini, “The Indexable Web is more than 11.5 billion

pages," 2005.
http://www.cs.uiowa.edu/~asignori/web-size/
8) 国立国会図書館,「諸外国の国立図書館等におけるデジタルアーカイブへの対応」, 2007.
http://www.mext.go.jp/b_menu/shingi/bunka/gijiroku/021/07073007/001.pdf
9) 国立国会図書館,「電子情報の長期的な保存と利用」
http://www.ndl.go.jp/jp/aboutus/preservation_02.html
・電子情報保存に係る研究調査報告書（平成15年）
・電子情報の長期保存とアクセス手段の確保のための調査報告書（平成16年，平成17年）
10) 国立国会図書館,「日本のWebサイトの網羅的収集，蓄積及び保存に関する調査報告書」, 2005.
http://www.ndl.go.jp/jp/aboutus/bulkresearch2005summary.html
11) NISO, "A Framework of Guidance for Building Good Digital Collections" (3rd Edition), 2007.
http://www.niso.org/framework/
12) 西尾 章治郎他,「情報の構造化と検索」, 岩波講座 マルチメディア情報学〈8〉, 岩波書店, 2000.
13) 鳥澤孝之,「日米における著作権法の図書館関係制限規定の見直しの動き」, カレントアウェアネス, No.289, CA1604, 2006.
http://current.ndl.go.jp/ca1604
14) 日本図書館協会 資料保存委員会,「りーふれっと資料保存」, 1999.
http://www.jla.or.jp/hozon/

Technical Problems in the Development of Digital Archives: Born-Digital Contents and Web Archives

KAWANO Hiroyuki

Abstract

In recent years, the density of recording media highly increases and the size of disk storage grows up rapidly, moreover the trends in electronic markets also change drastically. We have crisis of disappearing digital heritage due to various troubles of hardware and software. Actually, we will have many problems to browse and display digital contents within a few decades. All over the world, various organizations, such as libraries and museums, try to develop digital archive systems in order to preserve digitized contents and born-digital contents in the internet. In this paper, we introduce the digital and web archive services in several countries and argue technical problems to have long-term preservation of digital contents. We introduce technical terms such as METS, MODS, OAI-PMH and metadata schema, and also the essential software to collect, store, preserve and discover digital contents.