

デジタルアーカイブ・システム標準仕様書

国立公文書館

平成21年3月

目次

I デジタルアーカイブ・システム標準仕様

1 デジタルアーカイブ・システム標準仕様(標準仕様の考え方)	1
2 システム構築概要図	4

II デジタルアーカイブ・システムの構築について

1 デジタルアーカイブ・システムの基本的な考え方	5
1-1 システム導入の背景及び目的	5
1-2 システム導入の効果	6
1-3 システムの基本理念	7
1-4 システムの標準仕様における考え方	7
2 目録情報のデータ化	9
2-1 横断検索のために必要な最小限の目録情報の項目	9
2-2 記述方式、電子符号化	11
2-3 階層検索のために望ましいデータ項目、レベル	13
2-4 文字コード	13
3 目録情報のデータベース化	15
3-1 目録データの管理・登録機能	15
3-2 検索機能	18
3-3 業務機能	18
3-4 ユーザ・インターフェース	20
3-5 バックアップ機能	22
4 資料のデジタル画像化	23
4-1 画像データ	23
4-2 画像データの登録・管理機能	24
5 インターネットにおける情報提供	27
5-1 情報発信に必要な機能	27
5-2 横断検索機能	27
5-3 一括提供機能	28

6	ハードウェア	29
6-1	基本構成	29
6-2	ネットワーク	31
7	その他	32
7-1	構築作業	32
7-2	運用・保守	32

I デジタルアーカイブ・システム標準仕様

1 デジタルアーカイブ・システム標準仕様（標準仕様の考え方）

本項は、デジタルアーカイブ・システムの「標準仕様の考え方」について検討した結果を整理し、「デジタルアーカイブ・システム標準仕様」として取りまとめたものである。なお、検討の内容は、“II デジタルアーカイブ・システムの構築について”に取りまとめた。

表1 デジタルアーカイブ・システム標準仕様（標準仕様の考え方）

凡例 ◎：標準項目 ○：推奨項目

目録情報のデータ化	横断検索のために必要な最小限の目録情報の項目	◎ Dublin Core の要素を参考に、横断検索のために必要な最小限の目録データの項目として、「ID」「年代（作成年度）」「資料名」「作成者（部署名）」「備考等」の5項目を設定する。
	記述方式、電子符号化	◎ 目録情報の記述方式としてEAD/XML を採用する。（EAD/XML データの仕様は、EAD2002 を採用）
	階層検索のために望ましいデータ項目、レベル	◎ 公文書等に関する階層構成は、公文書等の移管元組織を反映したものとする。
	文字コード	◎ UCS-4 を表現できる文字コードとする。 ◎ 符号化形式はUTF-8 またはUTF-16 とする。 （以下、外字機能（推奨）について） ○ 外字は、ISO/IEC10036 に準拠する「文字図形番号」（情報処理学会試行標準IPSJ—TS0002：2004）を用いる。 ○ 外字を指定するときには、外字とその外字に代替可能なUCS 規格にある文字（代替文字）を併せて指定する。 ○ XML テキストデータでの外字記述方式はEGIX（Embedding Glyph Identifiers In XML Documents）とする。

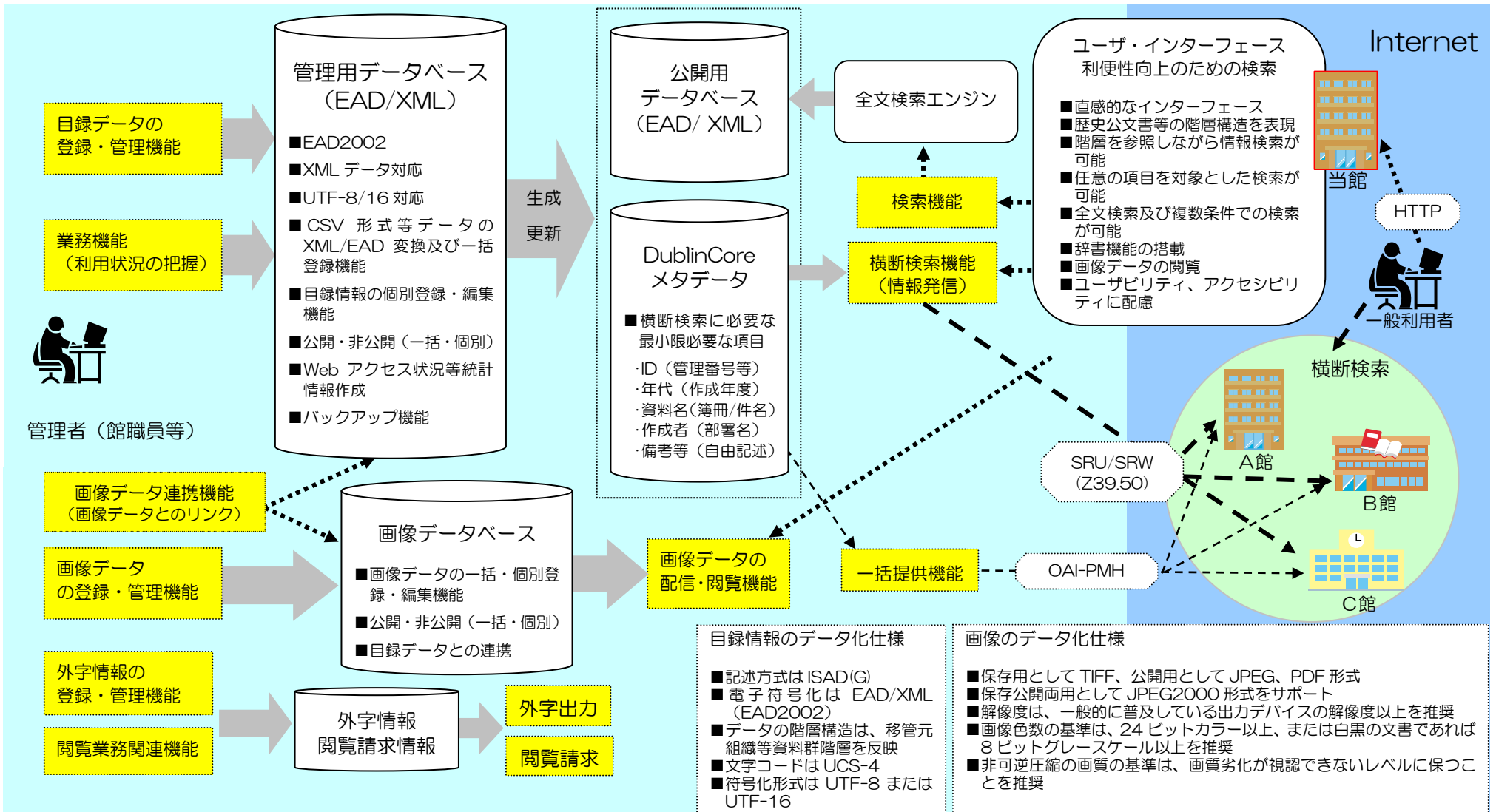
目録情報のデータベース化	目録データの管理・登録機能	目録データの移行	◎ CSV 形式等のデータをEAD/XML 形式に変換し、システムに登録・管理できる機能を設ける。
		データベース	◎ XML データを扱うことができるデータベースを採用する。
		検索エンジン	◎ XML データを検索できる全文検索エンジンを採用する。 ◎ UTF-8 またはUTF-16 を取り扱える全文検索エンジンを採用する。
		データの登録、編集、削除機能	◎ 目録データの登録、編集、削除機能を設ける。 ◎ 表計算ソフト等を用いて作成した目録データを、システムに一括登録が可能な機能を設ける。
		公開・非公開機能	◎ 個人情報の保護等へ配慮した機能を持つものとする。
	検索機能	利便性の向上のために必要な検索機能	◎ 複数の階層を指定して、その階層以下に属している目録データに限定して検索できるようにする。 ◎ EAD/XML の階層構造を階層毎に表示し、その階層を参照しながら情報を閲覧する検索方式を提供する。 ◎ 任意の項目を対象とした検索を可能とする。 ◎ 項目を限定しない全文検索ができ、これら複数の条件を組み合わせでの検索を可能とする。
	業務機能	利用状況の把握機能	◎ Web アクセス状況等の統計情報を作成し、利用状況を把握可能とする。
		閲覧業務関連機能等	○ 閲覧請求書印刷や閲覧者登録等の閲覧業務関連機能を推奨する。
ユーザ・インターフェース		◎ 階層構造が表現されるユーザ・インターフェースを採用する。	
バックアップ機能		◎ バックアップを取ることができる機能を有するものとする。	
資料のデジタル画像化	画像データ	画像形式	◎ 保存用としてTIFF、公開用としてJPEG、PDF を画像形式としてサポートする。 ○ 保存公開両用としてJPEG2000 を画像形式としてサポートすることを推奨する。
		解像度、色数等	○ 解像度の基準は、一般的に普及している出力デバイスの解像度以上を推奨する。 ○ 画像色数の基準は、24 ビットカラー以上、または白黒の文書であれば8 ビットグレースケール以上を推奨する。 ○ 非可逆圧縮の画質の基準は、画質劣化が視認できないレベルに保つことを推奨する。
		画像データの公開・非公開機能	◎ 個人情報の保護等へ配慮した機能を持つものとする。
	画像データの登録・管理機能	画像データ・目録データの連携機能	◎ 目録データの検索の結果、件名情報に該当する画像データが存在する場合には、画像データへのリンクを表示し、画像配信機能により閲覧可能とする。
		画像データの配信・閲覧機能	◎ 画像データは、特別なソフトウェアを必要とせず表示できる方式とする。

インターネット における情報提 供	情報発信に必要な機能		◎ HTTP (IETF RFC2616) にて情報発信を可能とする。
	横断検索機能	必要最低限のシステム対応メ タデータ	◎ 情報発信および横断検索はデータの特性を問わず検索ができるようDublin Core を採用する。 ◎ EAD/XML とDublin Core のクロスワークを定義する。
		横断検索プロトコル	◎ SRU/SRW にて情報発信できるものとする。
	一括提供機能		○ OAI-PMH での情報発信機能を推奨する。
ハードウェア	基本構成	◎ 各公文書館等の予算や規模に応じたシステムを導入するものとする。	
	ネットワーク	◎ ネットワーク上のセキュリティに配慮する。	
その他	構築作業	◎ 既存のシステムからのデータ移行に配慮したシステムとする。	
	運用・保守	◎ システムの設計に当たっては、運用・保守等に配慮したシステムとする。	

2 システム概要図

本項では、前述の「デジタルアーカイブ・システム標準仕様」で整理した項目を図化した。

図1 システム構築概要図



II デジタルアーカイブ・システムの構築について

1 デジタルアーカイブ・システムの基本的な考え方

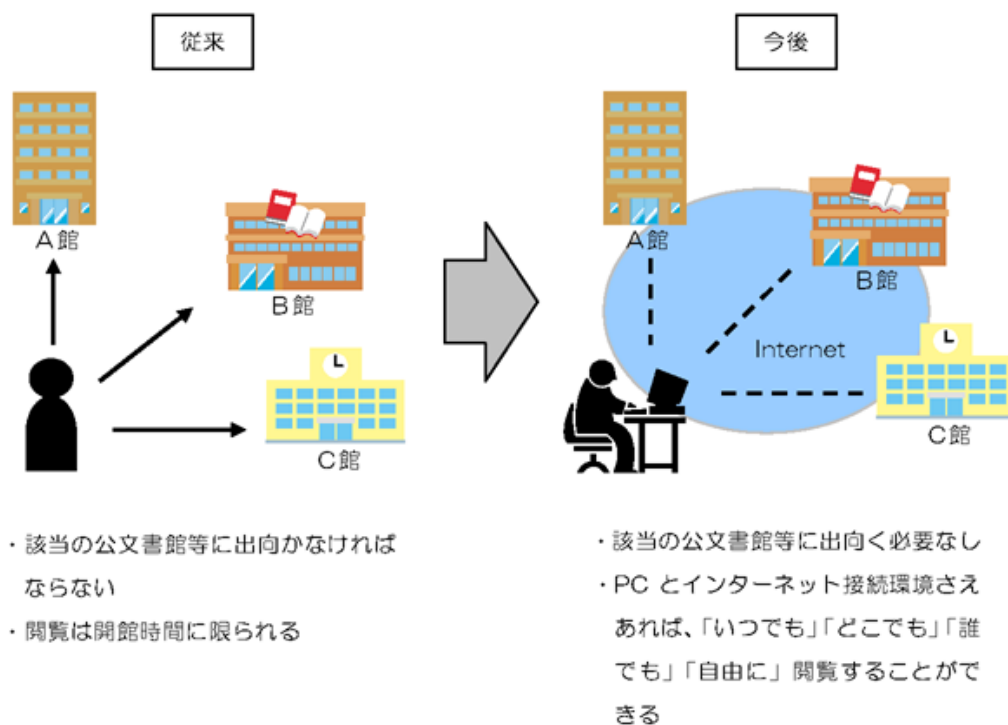
1-1 システム導入の背景及び目的

全国の公文書館等（以下、「公文書館等」という。）は、地方自治体の重要な歴史公文書等を適切に保存し、住民の利用に供している。

それら歴史公文書等は、地域はもとより、我が国の歩みを後世に伝える国民共有の財産であることから、情報通信技術を活用して「いつでも」「どこでも」「誰でも」「自由に」利用することが可能となるデジタルアーカイブ化を推進することが必要である。

デジタルアーカイブ・システムは、公文書館等が所蔵する公文書等の目録情報や利用者が利用し易い検索機能の提供、インターネットを通じて広く一般の利用者に閲覧可能とするなど、地域におけるデジタルアーカイブ化の推進を目的とする。

図2 デジタルアーカイブ・システムの導入イメージ



1-2 システム導入の効果

公文書館等は、デジタルアーカイブ・システムの標準仕様を踏まえつつ、地域に即したシステムを導入することにより、システム導入に向けて必要な調査や検討に要する工数の省力化とともに、システム構築経費の削減が期待される。

また、デジタルアーカイブ・システムの導入が順次進められることで、国や他地域の公文書館等を結びデジタルアーカイブ・ネットワークの構築が可能となる。これにより、インターネットを通じて所蔵する歴史公文書等の横断検索が実現され、各公文書館等における情報発信力の強化や全国規模での利活用が期待される。

図3 標準仕様を取り入れることによる経済的効果

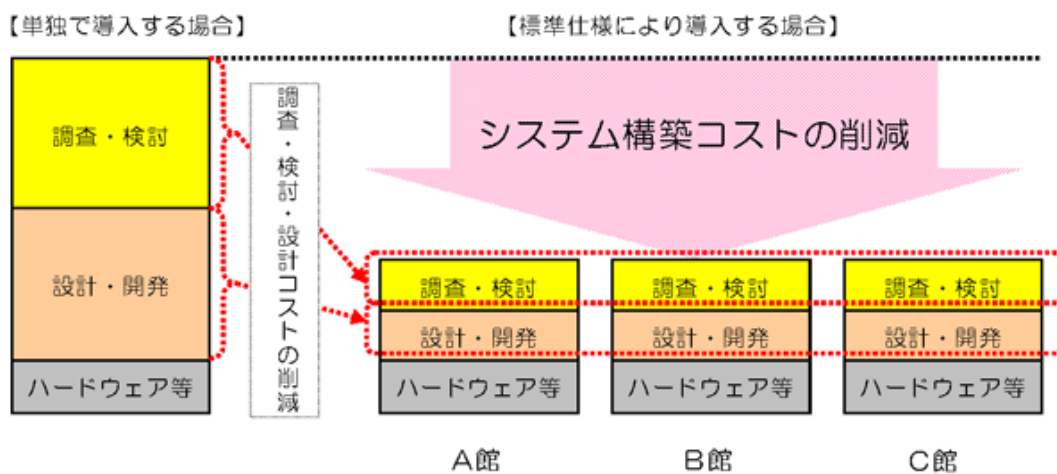
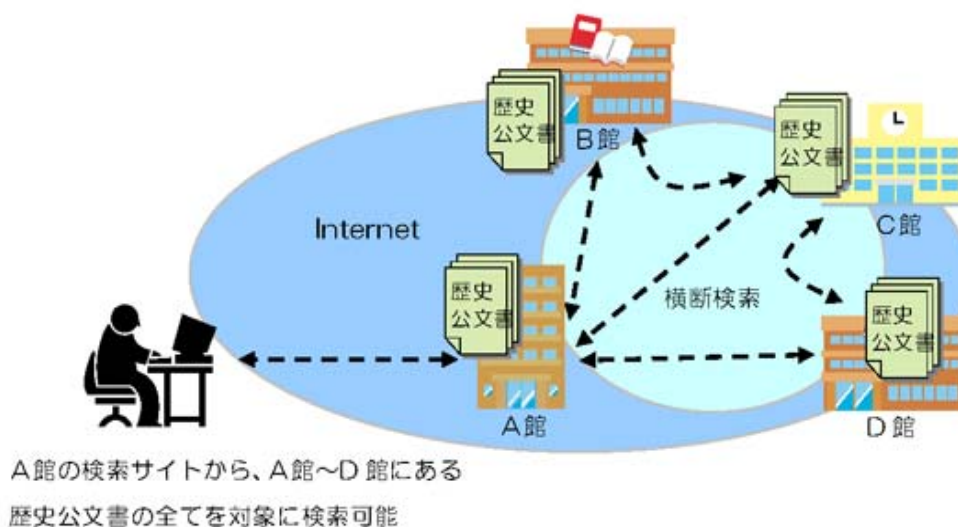


図4 インターネットを通じた歴史公文書等の横断検索のイメージ



1-3 システムの基本理念

公文書館等が所蔵する資料の内容や形態は非常に多岐にわたり、一般の住民に十分に理解されていない現状を考慮すると、いかに利用してもらえるようにすることが最大の課題である。

多くの公文書館等が現行のシステムで提供している簿冊名や件名のみ基礎的な目録情報では、一般の利用者が目的の資料にたどりつくことは困難であるということを念頭に置き、デジタルアーカイブ・システムは、コンピュータシステムとインターネットを利用して、利便性の高いシステムを実現し提供するものである。

これらを踏まえ、デジタルアーカイブ・システムは、導入の目的を達成するために、

- ① 全国の公文書館等の全てに対して提供でき、また現実的に利用可能なものとする。
- ② 公文書館等の個々の状況に応じ、柔軟に対応可能なものとする。
- ③ 公文書館等の現状を踏まえ、容易に導入でき、保守・運用が可能なものとする。
- ④ 「いつでも」「どこでも」「だれでも」「自由に」利用可能なサービスを実現するシステムとする。
を基本とする。

1-4 システムの標準仕様における考え方

デジタルアーカイブ・システムは、前述した機能要件を満たすために、システムの標準仕様における基本的な考え方を以下のように定める。

- ① システムが取り扱う資料の範囲

公文書を基本とするが、古文書等の資料についても対応可能な設計とする。

- ② 国際標準の採用

国際標準または事実上の標準の採用を基本とする。

- ③ 標準仕様の構成

(ア) 最小限必要と考えられる機能からなる以下の構成を基本機能仕様とする。

- 目録情報のデータ化
- 目録情報のデータベース化
- インターネットにおける情報提供
- 横断検索機能
- 画像データ連携機能

(イ) 公文書館等の要望や導入規模に合わせて、以下の機能を選択的に追加可能な推奨機能仕様とする。

- 外字機能

- ④ 公的なシステムとしての条件

公的機関に求められるガイドライン等に従い、以下の条件を踏まえたものとする。

- オープンなシステムもしくは標準的なシステムであること
- 情報セキュリティへの対応
- その他、国又は自治体が示す要件を満たすものであること

さらに、この基本的な考え方に基づく標準仕様項目を次に定める。

表2 デジタルアーカイブ・システムの標準仕様項目

目録情報のデータ化	横断検索のために必要な最小限の目録情報の項目	
	記述方式、電子符号化	
	階層検索のために望ましいデータ項目、レベル	
	文字コード	
目録情報のデータベース化	目録データの管理・登録機能	目録データの移行
		データベース
		検索エンジン
		データの登録、編集、削除機能
		公開・非公開機能
	検索機能	利便性の向上のために必要な検索機能
	業務機能	利用状況の把握機能
		閲覧業務関連機能等
ユーザ・インターフェース		
バックアップ機能		
資料のデジタル画像化	画像データ	画像形式
		解像度、色数等
	画像データの登録・管理機能	画像データの公開・非公開機能
		画像データ・目録データの連携機能
		画像データの配信・閲覧機能
インターネットにおける情報提供	情報発信に必要な機能	
	横断検索機能	必要最低限のシステム対応メタデータ
		横断検索プロトコル
	一括提供機能	
ハードウェア	基本構成	
	ネットワーク	
その他	構築作業	
	運用・保守	

2 目録情報のデータ化

2-1 横断検索のために必要な最小限の目録情報の項目

目録情報の横断的な検索は、現時点ではごく一部の公文書館等と国立公文書館との間に留まっている。しかし、所蔵資料の利活用をより一層促進するためには、各公文書館等のデジタルアーカイブ・システムのネットワーク化による横断検索を実現することが必要である。

EAD という標準的な目録データの表現方式を採用したとしても、細部における目録データの項目は、各公文書館等の事情により異なることも想定され、そのままでは横断検索において有効な結果を得ることは難しい。横断検索を実現するためには、目録データの項目を合わせる必要があるが、メタデータ記述項目の国際標準としては、15 の基本要素で構成されるDublin Core がある。

Dublin Core は、Web 上リソースのメタデータを記述するための定義であり、抽象化されたメタデータ項目であるために、あらゆる分野に適用することができるという特徴がある。従って、横断検索を含めた目録データの相互互換性を高めることが可能である。横断検索時の目録データの項目については、「ID」「年代（作成年度）」「資料名」「作成者（部署名）」「備考等」を挙げる公文書館等が多くあるとともに、検索行動に関する調査では、Dublin Core の基本要素のうち、主に利用されている検索項目は「Title」「Creator」「Subject」「Description」「Type」等であるとの結果が報告されている。

このことから、目録データの項目全てを用いて横断検索を行うのではなく、公文書館等が作成する目録データの中から、「ID」「年代（作成年度）」「資料名」「作成者（部署名）」「備考等」の5項目を横断検索に必要な最小限の項目とすることが考えられる。ただし、「ID」については、各公文書館等における資料特定のために用いる番号（所蔵資料番号等）であることから検索項目としては適さないため、表示項目としてのみ採用するものとする。

【標準仕様の考え方】

- Dublin Core の要素を参考に、横断検索のために必要な最小限の目録データの項目として、「ID」「年代（作成年度）」「資料名」「作成者（部署名）」「備考等」の5項目を設定する。

※参考1

http://www.jstage.jst.go.jp/article/jsik/17/1/15/_pdf/-char/ja/（H20.2.26）

検索行動調査に基づく検索エレメント設計に関する一考察（松村敦、古川沙希子、宇陀則彦）

表3 Dublin Core 要素一覧

要素名	ラベル	説明
(1)タイトル	Title	オブジェクトの名前
(2)著者あるいは作者	Creator	情報資源の内容に関して責任を持つ人または組織
(3)主題およびキーワード	Subject	情報資源に述べられたトピック
(4)内容記述	Description	アブストラクトやイメージデータの説明など内容に関する記述
(5)公開者（出版者）	Publisher	情報資源を現在の形態にしたもの（出版社、大学など）
(6)寄与者	Contributor	著者ではないが文書の内容の作成に関わった人または組織（編集者や翻訳者等）
(7)日付	Date	現在の形で利用できるようになった日付
(8)資源タイプ	Type	ホームページ、小説、詩、辞書といった情報資源の型
(9)形式	Format	Postscript ファイルやWindows 実行形式といった、情報資源のデータ形式
(10)資源識別子	Identifier	情報資源を一意に識別するための番号あるいは名前
(11)情報源（出处）	Source	情報資源の出所となった情報資源を一意に示す番号もしくは文字列
(12)言語	Language	情報資源の内容を記述している言語
(13)関係	Relation	他の情報資源との関連づけ
(14)対象範囲（空間的・時間的）	Coverage	地理的場所や時間的な内容に関する情報資源の特性
(15)権利管理	Rights	著作権記述などの権利に関する記述や利用条件に関する記述へのリンク(URL もしくは何らかのURI)

※参考2

http://www.dl.slis.tsukuba.ac.jp/DLjournal/No_14/1-sugimoto/1-sugimoto.html (H20.2.26) Dublin Core Metadata Element Set について — 現在の状況と利用例（杉本重雄）表1「Dublin Coreの15エレメント」より

2-2 記述方式、電子符号化

公文書館等の所蔵資料は、簿冊形式の公文書から絵図、巻物、古文書等、多種多様である。デジタルアーカイブ・システムでは、主に共通してみられる公文書や古文書等の目録情報が取り扱えることが求められる。目録情報の作成単位は、簿冊単位と件名単位の両方がみられることから、システムの目録データも、これに対応可能であることが望ましい。

目録情報の項目は、簿冊単位の場合では「簿冊情報」「部署情報」「年代情報」「選別等作業情報」「分類情報」、件名単位では「件名情報」「年代情報」「簿冊情報」「作成部署」「備考」「分類情報」が共通してみられる主な項目として捉えることもできるが、詳細は公文書館等によって相当の差がある状態である。しかし、8割弱の公文書館等が独自方式を採用している目録データの記録史料記述方式については、標準化する必要がある。

記録史料記述における国レベルでの標準には、DACS（Describing Archives: A Content Standard、米国）、RAD（Rule for Archival Description、カナダ）、MAD（Manual of Archival Description、英国）等がある。国際基準の記述方式としては、1992～96年に国際文書館評議会（ICA：International Council on Archives）が制定したISAD(G)（General International Standard Archival Description）がある。

ISAD(G)は記述に関する考え方で、これをホームページ等の作成に用いるマークアップ言語のXML（Extensible Markup Language）形式で表現した符号化記録史料記述がEAD（Encoded Archival Description）である。EADは、電子的検索手段の符号化における事実上の国際標準であるとともに、目録データを階層構造で表現することが可能である。

一方、一部の公文書館等では、図書館との複合施設として運用されているほか、所蔵資料の種類として図書等が存在する公文書館等もあった。図書目録の書誌記述には、国際標準であるISBD(G)（International Standard Bibliographic Description）のほか複数の方式が存在するが、国内ではMARC（Machine Readable Cataloging）が一般的に採用されているのが現状である。

また、Dublin Coreは基本的な15エレメントを定義することによってインターネット上の多様な情報資源に対応し、情報資源発見の効率化をはかるために定義されたコアメタデータである。Dublin Coreは単にシンプルなメタデータとして理解するのではなく、様々な分野に共通な要素、すなわち様々な分野でできるだけ共通の概念として認められた属性の要素の集まりとして定義されたコアメタデータとして理解すべきである。Dublin Coreが担う最も重要な役割は相互利用性（Interoperability）であり、Webコンテンツ、既存の各種メタデータなど多様な情報資源をDublin Coreをゲートウェイとして相互に結ぶことで、情報資源の発見が容易になる。

これらを踏まえ、デジタルアーカイブ・システムでは、公文書や古文書等の資料を対象とした記録史料の記述方式と、図書館との連携や所蔵する図書等への対応として図書目録の記述方式に対応するとともに、両者の相互利用を可能とすることが望ましい。

【標準仕様の考え方】

- 目録情報の記述方式としてEAD/XMLを採用する。（EAD/XMLデータの仕様は、EAD2002を採用）

※参考3

<http://www.loc.gov/ead/ead.html> (H20.2.26)

EAD: Encoded Archival Description

<http://www.ica.org/en/node/30000> (H20.2.26)

ISAD(G): General International Standard Archival Description, Second edition

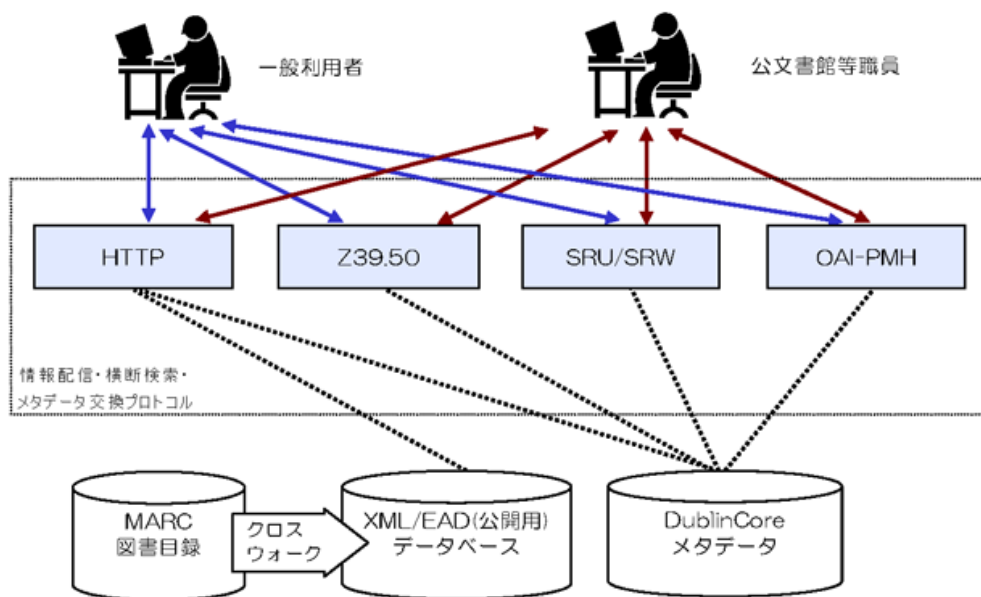
http://archives.nijl.ac.jp/DAS/projects/eadfa/20051020_gotoh.htm (H20.2.26)

EAD の概要と日本における動向：国文学研究資料館の事例紹介を中心に（五島敏芳）

http://www.nijl.ac.jp/~kiban-s/project/seika_pdf/2002/2002-012.pdf (H20.2.26)

「国際標準：記録史料記述の一般原則；ISAD(G)」とXML の利用（五島敏芳）

図5 EAD/XML、MARC、Dublin Core の関係



2-3 階層検索のために望ましいデータ項目、レベル

所蔵資料を検索するにあたり、全体の概要からどこに欲しい情報が存在するか、おおよその当たりを付けられるような階層構造を用いて、目録データを管理することが必要である。

資料の出所は、資料の種類や時代背景・集合としての目的や意味を包含していることが多いことから、公文書に関しては移管元組織（例えば、局→部→課→係（担当））での階層構成が適切と考えられる。

利用者に優先して提供されるべき階層情報が、最新の部署による階層情報であるとした場合、階層情報の適切な維持・管理によって、最新の部署階層による表現とすることが望ましい。また、階層の変更履歴をデータとして管理することにより、その変遷を辿ることができるような仕組みを準備することも考えられる。

【標準仕様の考え方】

- 公文書等に関する階層構成は、公文書等の移管元組織を反映したものとす。

2-4 文字コード

電子データとして目録情報を作成する際に使用している文字コードは、Shift-JIS が約半数を占めているが、文字コードの標準仕様を定める必要がある。

文字コードには、世界の主要な言語で使われる文字のほとんどを収録する国際文字集合である ISO/IEC10646-1 として定められた UCS (Universal multi-octet coded Character Set) がある。UCS には31 ビット (4バイト) の UCS-4 と、その部分集合である16 ビット (2バイト) の UCS-2 があるが、UCS-2 の場合、2 バイト表記では最大65536 文字しか収録できないため、中国語・日本語・韓国語で同じ意味や同じルーツの漢字は全て同じ文字とみなし、同じコードを割り当てる統合作業が行われ、適正に表現できないことがある。このことから、UCS-4 をカバーできる文字コードとすることが望ましい。

一方、UCS を用いたコンピュータ上で扱う際の符号化方式としては、UTF (Unicode Transformation Format) が使われるのが一般的である。代表的な UTF には、UTF-8とUTF-16 があり、それぞれ文字を表現する時の単位に違いがある。UTF-16 の場合、2 バイトで表現できる文字はそのまま2 バイトで表し、それ以降の文字は4 バイトで表す。このため、英数も日本語も全て2 バイトで表現される。UTF-8 の場合、英数は1 バイトで表現し、日本語は3 バイトで表現するようになっている。

これらを踏まえると、英数の割合が多い場合は UTF-8、日本語が多い場合は UTF-16 が良いことになるが、デジタルアーカイブ・システムでは、公文書館等が作成する目録データを考慮して、UTF-8 と UTF-16 を使い分けることが考えられる。

また、各公文書館等により取り扱う資料に差はあり、特に古文書等でみられる特殊な文字が多い場合には、UCS-4 に対応しない文字の扱いとして、外字機能を取り込むことを推奨する。なお、データベース化にあたっては、文字コードでは対応できない異体字・外字問題が発生する。これらは「文字図形識別情報」に基づき整理された「文字図形番号」を用いた情報交換が推奨されるが、既存のデータベースには異体字・外字部分に図形画像を張り込んで表示させるという方法が多くとられており、該当部分が検索出来ない。

情報アクセスの高機能化という観点から、一連のテキスト検索と同時にそこに含まれる画像が同時に検索出来ることが望ましい。

【標準仕様の考え方】

- UCS-4 を表現できる文字コードとする。
- 符号化形式はUTF-8 またはUTF-16 とする。
(以下、外字機能(推奨)について)
- 外字は、ISO/IEC10036 に準拠する「文字図形番号」(情報処理学会試行標準IPSJ—TS0002:2004)を用いる。
- 外字を指定するときには、外字とその外字に代替可能なUCS 規格にある文字(代替文字)を併せて指定する。
- XML テキストデータでの外字記述方式はEGiX (Embedding Glyph Identifiers In XML Documents) とする。

※参考4

http://www.jsa.or.jp/stdz/instac/syokai/h18jigyuhoukoku/h18annual-report/02_01.pdf
(H20.2.26)

ユビキタス社会を推進する情報基盤の標準化調査研究成果報告書(財団法人日本規格協会・情報技術標準化研究センター)

3 目録情報のデータベース化

3-1 目録データの管理・登録機能

(1) 目録データの移行

現況の公文書館等における目録データの管理は、EXCEL 等の表計算ソフトやカード型データベース、小規模なRDBMS 等により行われている例が多い。これらの環境から、デジタルアーカイブ・システムへスムーズに移行可能とすることが必要である。

また、一部の公文書館等では、既に電子公文書の受入れ（移管）に対応したシステムを導入している。情報化の推進によって、電子文書管理システム等を導入する自治体は、今後増加していくものと予想され、公文書館等もその対応が必要となってくる。しかし、デジタルアーカイブ・システムにおける電子公文書のデータそのものへの対応については、現時点では電子文書管理システム等が普及の初期段階にあること、システムの連動を構築することは、デジタルアーカイブ・システムの構築費用の増大及び電子文書管理システム等側にも改造が必要なることが想定される。

これらを踏まえ、現行のシステムからのデータ移行や、電子文書管理システム等とのデータ連携については、システム依存度が低く、多くのアプリケーションでサポートされているCSV（Comma Separated Values）形式等のデータをEAD/XML 形式に変換し、システムに登録・管理できる機能を設けることが望ましい。

【標準仕様の考え方】

- CSV 形式等のデータをEAD/XML 形式に変換し、システムに登録・管理できる機能を設ける。

(2) データベース

データベースシステムを運用している公文書館等は、都道府県で6割強あった。市区町では、EXCEL 等の表計算ソフトを用いて目録情報を管理している例も半数みられる。現状では、目録データの管理・登録に使用しているシステムは、小規模なものから大規模なシステムまで、また、所蔵資料の種類によって複数のデータベースを運用している例など様々である。このような現状であるが、デジタルアーカイブ・システムでは、XML のデータ形式を利用することから、XML データを扱うことができるデータベースが望ましい。

なお、目録データの登録や管理にあたっては、ある条件に合致するデータを一括で取り扱うなど、データの編集や抽出を効率的に行うことが求められる。この点からリレーショナル・データベースのようなデータベースを導入することも有効であると考えられる。

【標準仕様の考え方】

- XML データを扱うことができるデータベースを採用する。

(3) 検索エンジン

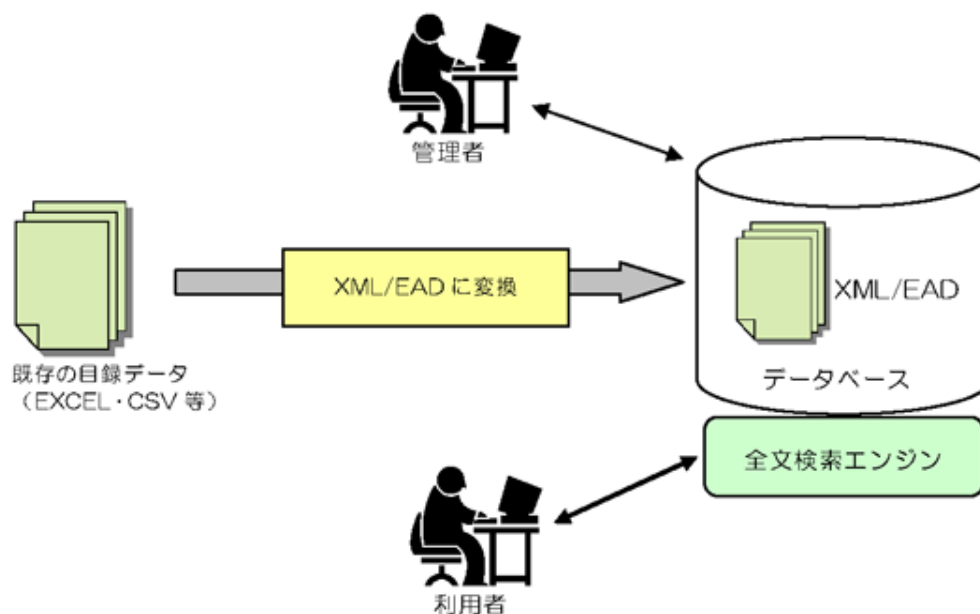
デジタルアーカイブ・システムは、2-4 文字コードにおいて触れたとおり、UTF-8またはUTF-16 を取り扱うとともに、データベースへはXML データとして登録されることから、この両者に対応した検索エンジンが必要と考える。

また、検索は、目録の項目ごとに検索を行うことはもとより、公文書等で用いられる用語や目録の項目になじみのない一般利用者が公文書等の検索を行うことを想定すれば、与えられた検索ワードを目録の項目にとらわれず検索することができる全文検索が可能な検索エンジンを採用することが、必要不可欠であると考えられる。

【標準仕様の考え方】

- XML データを検索できる全文検索エンジンを採用する。
- UTF-8 またはUTF-16 を取り扱える全文検索エンジンを採用する。

図6 目録データの移行の流れとデータ利用のイメージ



(4) データの登録、編集、削除機能

資料の受入れ（移管）の頻度は、年1回が4割弱で、その他が7割を占めている。その他には、随時のほか、公文書は年1回・古文書や歴史資料等は随時、保存年限によって公文書の受入れ時期をずらしている公文書館等もある。このことから、デジタルアーカイブ・システムへの目録データの登録は、期間を限定せず随時行えることが求められる。なお、データの登録にあたっては、資料の受入れ時に全ての目録データの項目を入力することは難しいことも想定し、必須項目のみ先行入力し、後から情報を追記可能とすることが必要である。

また、データベースシステムやACCESS等のデータベースソフトを使用している多くの公文書館等では、誰もが比較的簡易に使用できるEXCEL等で目録データを作成した後、システムへ登録している例がみられた。よって、デジタルアーカイブ・システムでは、1件毎の目録情報の登録機能とともに、EXCEL等でシステムの目録情報の項目に合せ作成されたCSV形式等のデータを一括して登録可能とする機能が必要である。

【標準仕様の考え方】

- 目録データの登録、編集、削除機能を設ける。
- 表計算ソフト等を用いて作成した目録データを、システムに一括登録が可能な機能を設ける。

(5) 公開・非公開機能

プライバシー保護への意識の高まりを受け、所蔵資料の公開にあたっては、個人情報の保護等に留意することが必要である。公文書館等では、個人情報の保護に関する規定等に基づき、資料の公開または非公開の判断が求められる。

これらを踏まえ、デジタルアーカイブ・システムにおいても、目録データに公開・非公開の項目を設け、非公開の場合、公開制限をかける仕組みが必要である。

【標準仕様の考え方】

- 個人情報の保護等へ配慮した機能を持つものとする。

3-2 検索機能

(1) 利便性の向上のために必要な検索機能

データベースシステムを運用している公文書館等では、キーワード、文書種別、文書分類等を用いた検索機能を装備している例が多くある。システムの更新（新規導入）に関する意向でも、任意の文字列による検索や階層検索の実装など、検索機能の充実を挙げる意見が多数みられた。このことから、デジタルアーカイブ・システムでは、任意の項目を対象とした検索、項目を限定しない任意の文字列による検索、これら複数の条件を組み合わせた検索など、多様な手段による検索が可能であることが望ましい。

また、資料の階層的な参照ができることや、特定の階層以下に限定した検索を可能とすることで利便性が高まるものと考えられる。検索結果に関連する資料は、階層を辿ることで容易に探し出すことができるため、該当資料がどの階層に属しているのか把握できるようにすることも重要と考えられる。

一方で検索機能を使用する際に、利用者が公文書等で用いられる用語をキーワードとしての確に入力できるとは限らないため、同意語彙の辞書を作成し、検索時に利用できるようにすることでより利便性の向上を図ることができる。

また、辞書に関連する機能として、利用者が辞書を使用する・しないを選択できる、利用者が検索に利用した語彙を把握できるようにし、利用者の混乱を避ける等の機能が考えられる。

【標準仕様の考え方】

- 複数の階層を指定して、その階層以下に属している目録データに限定して検索できるようにする。
- EAD/XML の階層構造を階層毎に表示し、その階層を参照しながら情報を閲覧する検索方式を提供する。
- 任意の項目を対象とした検索を可能とする。
- 項目を限定しない全文検索ができ、これら複数の条件を組み合わせでの検索を可能とする。

3-3 業務機能

(1) 利用状況の把握機能

頻度の高い目録情報の充実、辞書機能のメンテナンス、システム資源の見直し等を考慮して、システムの利用状況を把握する機能が必要であると考えられる。これにより、来館者数のみならず、インターネットを通じたデジタルアーカイブ・システムの利用件数も、公文書館等の利用状況として、年報や各種の利用実績の指標として活用することが想定される。

【標準仕様の考え方】

- Web アクセス状況等の統計情報を作成し、利用状況を把握可能とする。

(2) 閲覧業務関連機能等

公文書館等において原資料を閲覧する際に提出する閲覧請求書を検索結果より印刷することができる機能や閲覧者登録等をシステム上で取り扱うことができる機能を設けることなども可能であると考えられる。これら閲覧業務関連機能等を推奨する。

【標準仕様の考え方】

- 閲覧請求書印刷や閲覧者登録等の閲覧業務関連機能を推奨する。

3-4 ユーザ・インターフェース

デジタルアーカイブ・システムの目的の一つは、公文書館等が所蔵する公文書等の目録情報を一般の利用者に閲覧可能とすることである。従って、システムのユーザ・インターフェースは、誰もが使い易いことが必要である。

この点において、国立公文書館のデジタルアーカイブ・システムは、EAD で表現される階層情報をツリー型の表示により常に参照可能であるなど、利便性の高いシステムを実現したものである。

このようなツリー型の階層情報表示は、従来のキーワードによる目録情報の検索方法に比べて、登録されているデータの前提知識が無い利用者でも、階層をマウス操作により辿っていくだけで資料の検索を行うことができる。

ユーザ・インターフェースは、公文書館等ごとに実情に合わせたものを作成することによって使いやすいシステムを目指すべきであるが、上記のような階層構造を表現する方法（ツリー型など）に関しても留意すべきである。

【標準仕様の考え方】

- 階層構造が表現されるユーザ・インターフェースを採用する。

図7 階層情報をツリー型で表現している事例（国立公文書館）

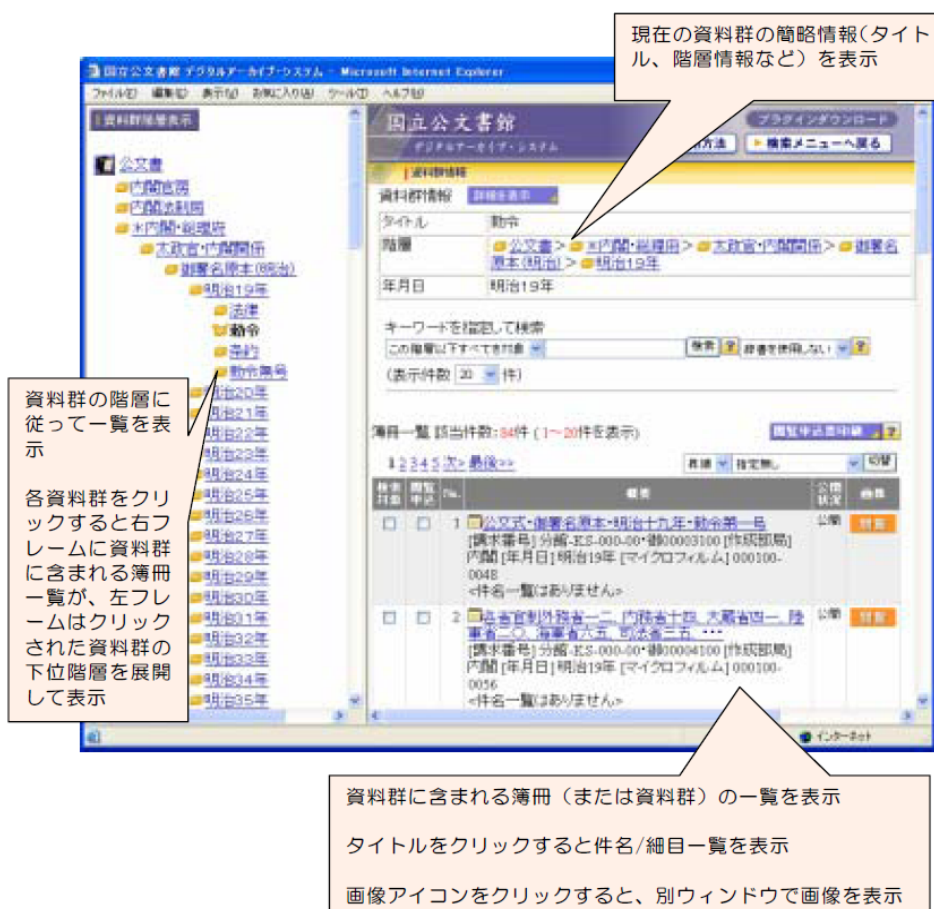
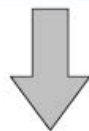
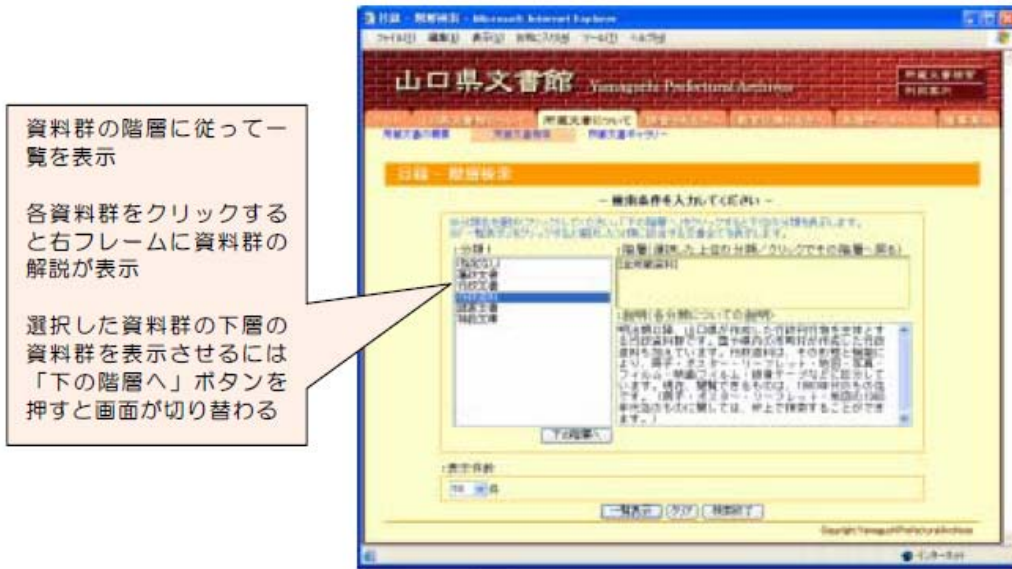


図8 階層情報をツリー型以外で表現している事例（山口県文書館）



資料群を選択していくと最終的に簿冊の一覧表示画面へ画面が切り替わる



3-5 バックアップ機能

システム障害へのリカバリにおいては、データの保存、保護が重要になる。一般的にも、システムの登録データは、一定期間毎に全体バックアップを行うことが重要とされる。

このことから、デジタルアーカイブ・システムにおける全体バックアップは、地震等の天災を考慮して、保管場所を分散できることも考慮すべきである。

【標準仕様の考え方】

- バックアップを取ることができる機能を有するものとする。

4 資料のデジタル画像化

4-1 画像データ

(1) 画像形式

デジタル画像化に取り組んでいる公文書館等で採用している画像形式は、汎用性や保管用としてJPEG が最も多く、次に、汎用性、さらには長期継続運用や利用者が操作しやすい等の理由から、TIFF がみられた。その他、RAW（無圧縮を選択でき、画像劣化が避けられるため）、ZOOMA（高解像度のデータ利用のため）、PDF、GIF を採用している例もある。目的別を理由に、複数の形式を併用している公文書館等もみられる。

事実上標準である形式及び国際標準化された形式としては、TIFF (Tagged Image File Format)、BMP (Bit MaP)、PNG (Portable Network Graphics)、JPEG (Joint Photographic Experts Group)、PDF (Portable Document Format)、JPEG2000の画像形式がある。このうち、TIFF は画像データの保存形式として用いられる例は多いが、公開を前提とした形式としては向いていない。PNG は可逆圧縮方式であるため、圧縮率が低くデータサイズが大きい（JPEG の約5～10 倍）、また、BMP は圧縮されていないため公開向けではない。JPEG2000 は、作成コストがかかることが難点ではあるが、高精細であり、画像の詳細部分を閲覧することが可能である。

これらを踏まえ、デジタルアーカイブ・システムは、既存資産の活用の観点から、保存用としてTIFF、公開用としてJPEG、PDF、また、保存公開の両用としてJPEG2000を画像形式とすることが望ましい。

【標準仕様の考え方】

- 保存用としてTIFF、公開用としてJPEG、PDF を画像形式としてサポートする。
- 保存公開両用としてJPEG2000 を画像形式としてサポートすることを推奨する。

(2) 解像度、色数等

画像データは、マイクロフィルムをソースとして作成される場合が多く、その解像度はマイクロフィルムによる制限を受ける。また、拡大印刷や増色処理などの特別な用途を除いては、出力デバイスよりも大きな解像度を持たせることに意義は認められない。マイクロフィルムの解像度は、現時点で普及している出力デバイスの解像度と比較して十分に高いと言えるため、解像度の基準は、一般に普及している出力デバイスの解像度以上（例えば、閲覧用120dpi 以上、漢字識別400dpi 以上、印刷用600dpi 以上または1200 dpi 以上）とすることが望ましい。

また、一般的な表示デバイスで正確に表現できる階調は256 であることから、グレースケールデータの階調もこれを上限として、ピクセルあたり8 ビットとすることが望ましい。

一方で、非可逆圧縮における画質の基準についても留意する必要がある。

【標準仕様の考え方】

- 解像度の基準は、一般的に普及している出力デバイスの解像度以上を推奨する。
- 画像色数の基準は、24 ビットカラー以上、または白黒の文書であれば8 ビットグレースケール以上を推奨する。
- 非可逆圧縮の画質の基準は、画質劣化が視認できないレベルに保つことを推奨する。

4-2 画像データの登録・管理機能

(1) 画像データの公開・非公開機能

画像データも目録情報と同様に個人情報の保護等が必要である。従って、デジタルアーカイブ・システムでは、画像データに公開・非公開の項目を設け、非公開の場合、公開制限をかける仕組みが必要である。

【標準仕様の考え方】

- 個人情報の保護等へ配慮した機能を持つものとする。

(2) 画像データ・目録データの連携機能

利用者の利便性を向上する観点から、目録情報を検索した際に関連する画像データが閲覧できることなど、デジタルアーカイブ・システムに画像データと目録情報を連携する機能を装備することが考えられる。

【標準仕様の考え方】

- 目録データの検索の結果、件名情報に該当する画像データが存在する場合には、画像データへのリンクを表示し、画像配信機能により閲覧可能とする。

図9 画像登録イメージ

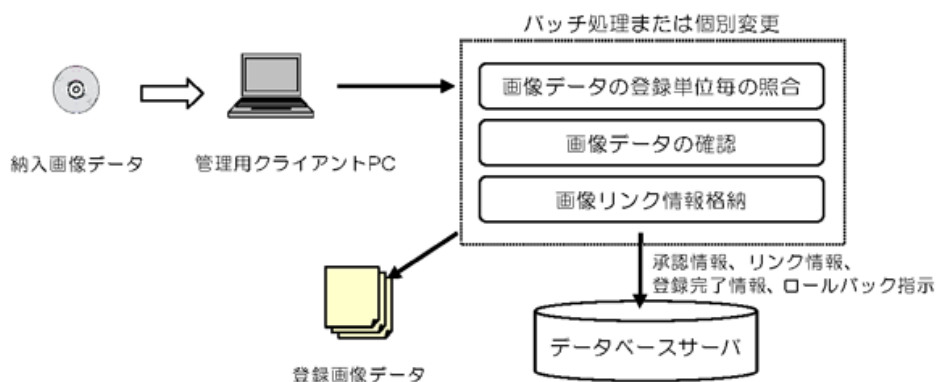
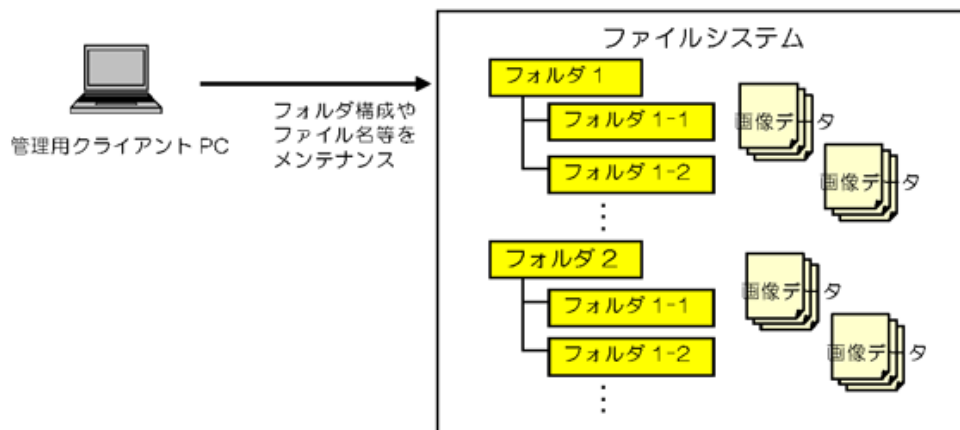


図10 画像データ保管に関するファイルシステム管理イメージ



(3) 画像データの配信・閲覧機能

目録の検索結果から連携して該当の資料画像が閲覧できることは、利用者にとっての利便性が増す。

また、誰もが画像データを参照できるよう、特殊なソフトウェアを必要とせず、WEBブラウザで画像を参照できることは必須要件と考える。WEB ブラウザから呼び出すプラグインで画像データを参照できることも想定されるが、この場合、様々なWEB 上のサイトにあるようなプラグインソフトウェアであることが重要である。

なお、JPEG2000 については、特別なプラグインソフトウェアが必要であるものの、国際標準であること、また前述のとおり高精細であり、画像の詳細部分を閲覧することが可能であるという観点から、国立国会図書館、国立公文書館でも採用されている。このため、推奨の対象とする。

【標準仕様の考え方】

- 画像データの配信・閲覧機能は、特別なソフトウェアを必要とせず表示できる方式とする。

5 インターネットにおける情報提供

5-1 情報発信に必要な機能

平成20年1月現在、ホームページを公開している公文書館等は47館あり、うち目録情報を提供しているのは23館、画像データを公開しているのは21館である。地理的・時間的な制約のない情報提供を実現するためには、国際規格の情報配信用プロトコルであるHTTP（Hypertext Transfer Protocol）（IETF RFC2616）を利用したインターネットによる情報発信を行い、WEBブラウザから資料の検索や参照を可能とする機能が必要である。

【標準仕様の考え方】

- HTTP（IETF RFC2616）にて情報発信を可能とする。

5-2 横断検索機能

（1）必要最小限のシステム対応メタデータ

前記したように、各公文書館等の目録データの横断検索を実現するためには、メタデータ記述項目における国際標準であるDublin Coreの要素を参考に、横断検索のために必要な最小限の目録データの項目を設定する必要があるが、その際、EAD/XMLとDublin Coreのクロスワークを定義することが必要となる。

【標準仕様の考え方】

- 情報発信および横断検索はデータの特性を問わず検索ができるようDublin Coreを採用する。
- EAD/XMLとDublin Coreのクロスワークを定義する。

（2）横断検索プロトコル

横断検索プロトコルとしては、国際標準のZ39.50、またZ39.50の後継プロトコルとしてSRU/SRW（ZING Search/Retrieve via URL/Search/Retrieve WebService）がある。SRU/SRWは国内でも普及が進んでおり、次世代の横断検索プロトコルとして期待されている。

Z39.50は、規格の策定から時間が経過しているものの、唯一の国際標準プロトコルであり、Z39.50およびSRU/SRWの双方を実装することが望ましい。

【標準仕様の考え方】

- SRU/SRWにて情報発信できるものとする。

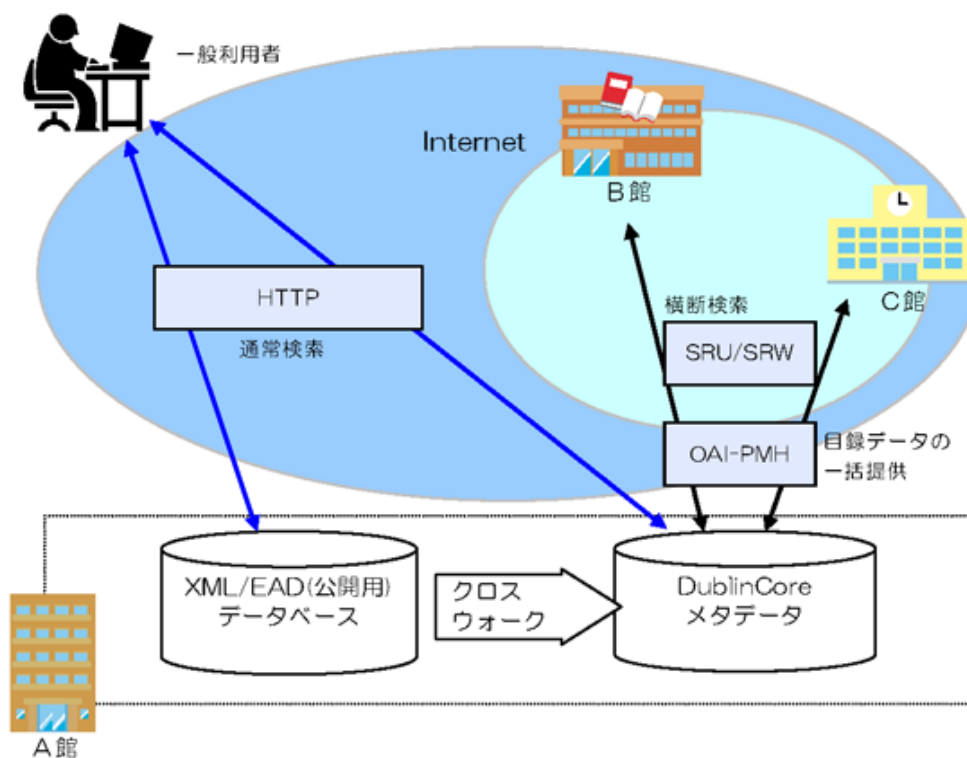
5-3 一括提供機能

一般利用者向けのHTTP による情報提供だけでなく、利用者の要求に対して該当するメタデータそのものを提供（あるいは受信）する仕組みも有効と考えられる。このようなメタデータ交換のための国際標準プロトコルとしては、OAI-PMH（Open Archives Initiative Protocol for Metadata Harvesting）が存在し、国内外でも学術機関を中心に、OAI-PMH に則ったメタデータ提供機能の実装が増えている。現時点においては、広く普及していないプロトコルであるが、今後、公文書館等においても他機関との連携を視野に入れ、OAI-PMH による情報提供を想定することは意義のあることと考えられる。

【標準仕様の考え方】

- OAI-PMH での情報発信機能を推奨する。

図11 インターネットにおける情報提供イメージ



6 ハードウェア

6-1 基本構成

データベースシステム等の更新や新規導入の予定が無い公文書館等は、全体7割強を占めている。その理由は、ここ数年内に更新や新規導入した例、現行のデータベースシステム等の機能に対して特に課題がない例を除くと、自治体における厳しい財政状況により、予算確保が困難であることを理由に挙げている例が多数ある。一方、更新や導入の予定がある公文書館等が想定している予算規模は、100万円以下から1,000万円程度までかなり幅広い意向がみられる。

これらを踏まえると、デジタルアーカイブ・システムは、各公文書館等の予算に応じた導入しやすい規模であることが求められ、図12~14のような構成例が考えられる。

【標準仕様の考え方】

- 各公文書館等の予算や規模に応じたシステムを導入するものとする。

表4 A 市現行システム概要

項目	内容	備考
導入年月	平成19年4月	
リース料	年間約400万円（サーバ1台、PC2台、プログラム等、ネットワーク機器は別途）	初年度費用（以降毎年同額の支払いが発生）
保守・運用関係費用	年間約150万円	
システム機能概要	【目録データ登録関連】 目録データ登録、目録データ管理（編集、削除） 【目録データ配信関連】 検索機能 【業務管理関連】 閲覧業務関連、利用状況把握、外字機能 【画像登録・配信関連】 画像登録、画像配信、画像閲覧	

*現地調査

図12 基本的な構成の例（サーバ1台による構成）

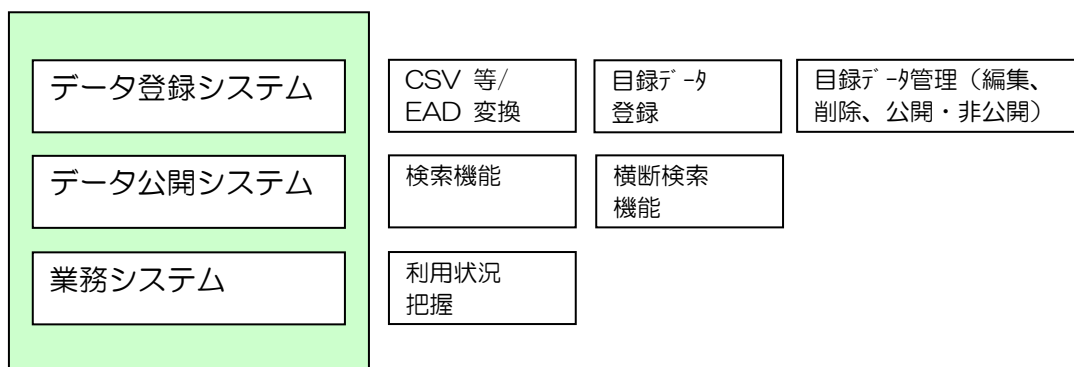


図13 標準的な構成の例（サーバ1台による構成）

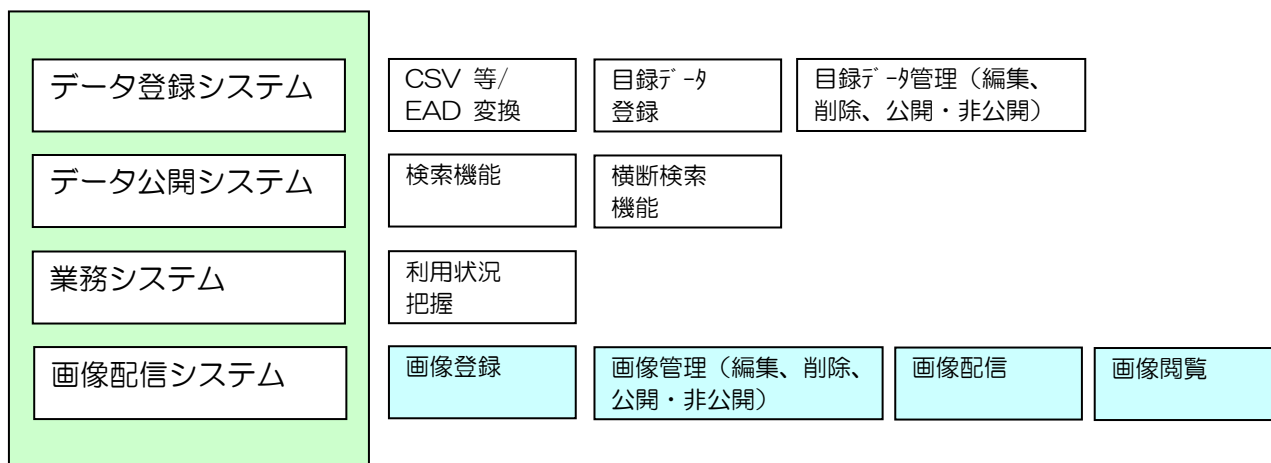
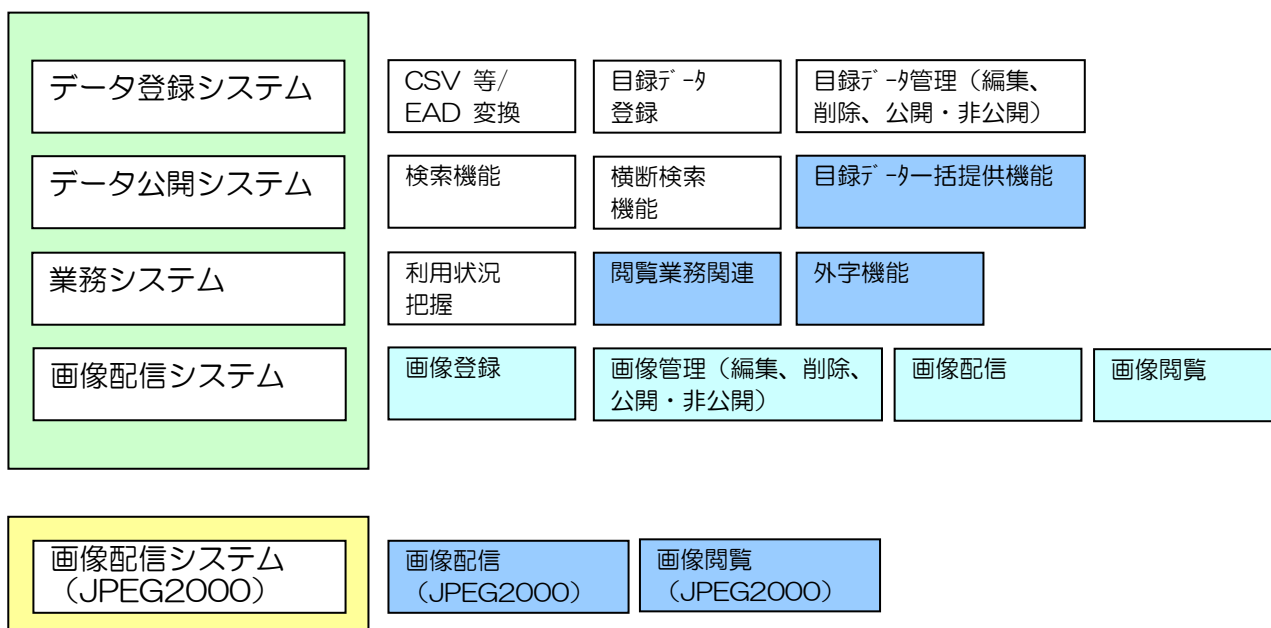


図14 発展的な構成の例（サーバ2台による構成）



6-2 ネットワーク

公文書館等においては、現状、セキュリティ確保のために物理的にデータベースシステム等と自治体の庁内LAN との接続を行っていない例や、ホームページ上で目録情報を提供するためのサーバへのデータ転送の際に、メディアを介して行うなどの工夫もみられた。

デジタルアーカイブ・システムは、ネットワークへの接続が必須であるとともに、その対象とするデータは、住民共有の財産であり、個人情報等も多く含まれることから、外部からの不正な通信を遮断する必要がある。そのためにも、ネットワーク上にファイアウォールを導入する必要がある。

【標準仕様の考え方】

- ネットワーク上のセキュリティに配慮する。

7 その他

7-1 構築作業

デジタルアーカイブ・システムの構築にあたっては、既存のネットワーク環境、電源設備、目録データ、画像データの状況により、調達範囲は流動的になると考えられる。従って、調達範囲については、設計、システム構築、試験、教育等を範囲とすることが考えられる。

また、調達範囲外のものとして、インターネット接続のための設置工事及び回線接続等の費用、空調設備及び間仕切り工事が想定されるほか、目録データ作成、画像データ作成、辞書データ作成など、データ整備に係る費用も想定される。なお、既存のシステムが存在する場合は、データ移行に関しても特段の配慮が必要となる。

【標準仕様の考え方】

- 既存のシステムからのデータ移行に配慮したシステムとする。

7-2 運用・保守

ホームページやシステム関連の業務を担当している職員数は、6割弱の公文書館等が1人以下、2～3人が約4割という状況である。システムの規模が大きくなると、運用面における負荷も増えることが想定される。よって、円滑なデジタルアーカイブ・システムの運用を行うためには、システムの運用や保守に関する仕様を定め、外部委託することにより効率化を図ることが必要である。

システムの運用・保守の具体的な内容については、公文書館等の状況に応じて、運用・保守の範囲を決定する必要があるが、公文書館等職員に対するシステムの利用及び運用に当たっての具体的操作の説明をはじめ、障害時における対処、公文書館等職員からの問い合わせへの対応等が考えられる。

【標準仕様の考え方】

- システムの設計に当たっては、運用・保守等に配慮したシステムとする。