琉球政府文書デジタル・アーカイブズ推進事業における資料画像の長期保存及び公開システムに関するレポート

大田　文子[†]

はじめに

「琉球政府文書デジタル･アーカイブズ推進事業」（以下「本事業」という。）は､沖縄県公文書館（以下「公文書館」という。）が所蔵する琉球政府が作成・収受した琉球政府文書のうち、13 万点をデジタル化し、インターネット公開する計画であり、沖縄県が 2013 年（平成 25）度より実施し、事業の終了は 2021 年（平成 33）度を予定している。公文書館の指定管理者である公益財団法人沖縄県文化振興会は、2016 年（平成 28）度に本事業のインターネット公開にかかる業務を受託し、インターネット閲覧者（以下「ユーザー」という。）の利便性向上のため、沖縄県が 2015 年（平成 27）度に開発した「琉球政府文書デジタルアーカイブ」公開システム（以下「公開システム」という。）の改修を行い、2018 年（平成 30）度 10 月末現在、13,500 点の資料をウェブ公開（以下「公開」という。）している。

本稿は、本事業を開始した 2013 年（平成 25）度から 2015 年（平成 27）度まで（以下「初期」という。）と、2016 年（平成 28）度から 2018 年（平成 30）度現在（以下「中期」という。）の資料の撮影及び画像データの作成仕様とその長期保存、そして、デジタル化資料の公開システムについて、変遷とその経緯を報告する。

1. 資料のデジタル化

1-1. 原資料の撮影

本事業における原資料の撮影は、基本的に綴じられている資料は綴じたまま撮影しており、撮影機材は事業当初より一貫して A3 サイズ対応のオーバヘッド式ブックスキャナを使用している。撮影仕様は表 1 に示す。

[†] おおた　あやこ　公益財団法人沖縄県文化振興会　公文書管理課　公文書専門員

表 1　原資料の撮影仕様

<table>
<tr><th rowspan="2"></th><th colspan="3">初期</th><th colspan="3">中期</th></tr>
<tr><th>2013</th><th>2014</th><th>2015</th><th>2016</th><th>2017</th><th>2018</th></tr>
<tr><td>撮影機材</td><td colspan="3">オーバーヘッド式ブックスキャナ
・A3サイズ対応</td><td colspan="3">オーバーヘッド式ブックスキャナ
・A3サイズ対応
・大判対応（A1サイズ）</td></tr>
<tr><td>撮影解像度</td><td colspan="3">350dpi以上</td><td colspan="2">300dpi</td><td>300dpi以上</td></tr>
<tr><td>表表紙</td><td colspan="3">資料コードが見える面を表表紙とする。</td><td colspan="3">資料コードが見える面を表表紙とする。
表紙にカラーチャート及びスケールを入れて撮影。</td></tr>
<tr><td rowspan="6">撮影対象文書</td><td colspan="6">見開き状態で左右両ページを撮影できる場合、両ページを1コマとして撮影。</td></tr>
<tr><td colspan="3">白紙のページは撮影しない。</td><td colspan="3">白紙のページも撮影する。</td></tr>
<tr><td colspan="3">大判文書（A3サイズ対応スキャナで文書の四隅が撮影ができない、かつ、2分割を超える分割撮影が必要な文書）が少ない資料は、資料は撮影対象とするが、大判文書のみ撮影しない（大判対応スキャナ導入年度後に一部分撮影する）。</td><td colspan="3" rowspan="2">大判文書（A3サイズ対応スキャナで文書の四隅が撮影ができない、かつ、2分割を超える分割撮影が必要な文書）は大判対応スキャナで撮影する。</td></tr>
<tr><td colspan="3">A3サイズ対応スキャナで文書の四隅が収まらないときで、撮影できない範囲に文字情報がない場合は、見切れていても分割しない。</td></tr>
<tr><td colspan="3">綴じられていて開けない資料は綴じたまま撮影する。</td><td colspan="3">綴じられていて開けない資料がある場合は、綴じを外す、または緩めて撮影する。</td></tr>
<tr><td colspan="3">利用制限情報である袋掛け及び付箋は外さず、該当文書の撮影はしない。</td><td colspan="3">利用制限情報である袋掛け及び付箋を取り除きすべて撮影する。</td></tr>
<tr><td>ノド部分の撮影</td><td colspan="3">可読できればよし。</td><td colspan="3">綴じを外す、または緩めて撮影する。</td></tr>
<tr><td>同一資料</td><td colspan="2">様式のみで同一資料が続く場合は同一資料の一部目を撮影し、その後は撮影しない。</td><td colspan="4">すべて撮影する。</td></tr>
</table>

事業の初期においては、利用制限情報判定済みの頁の撮影を省略し、代わりに「個人情報等があるため閲覧できません」と記したターゲットを差し込んでいた（図 1）。ここでいう「利用制限情報判定済み」とは、沖縄県公文書館管理規則第 4 条により利用が制限される文書について、すでに袋掛け等によるマスキング処置がなされたものをいう。

さらに大判文書の撮影も省略し、また、綴じを外さずに撮影する方針だったため、ノドの部分が読めないなどの不具合もあった。このため、初期に作成された画像データは資料の複製物としては不完全なものにならざるをえなかった。

しかし、中期以降はこの点を改善した（図 2）。2016 年（平成 28）度からは、必要な資料については綴じを外し、すべての頁の撮影を行い、利用制限情報部分のみを墨消しする「部分マスキング」を採用し、公開している。初期に撮影を省略した頁については、今後補っていく必要がある。

図 1　初期の公開画像

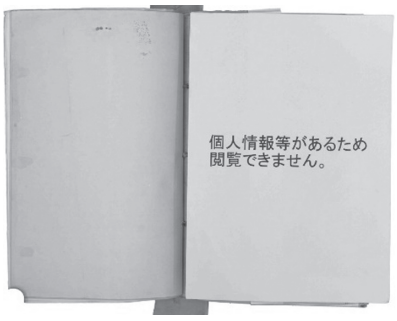
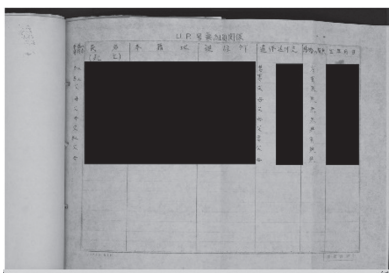


図 2　中期の公開画像



1-2. 画像データ作成

撮影後に作成する画像データの仕様にも、初期と中期では違いがある（表 2)。初期は、撮影したそのままの画像データである RAW データを閲覧に最適な圧縮率で圧縮した Jpeg 格納マルチ PDF（以下「PDF」という。）に変換していた。しかし、中期は RAW データをアプリケーションソフトに依存せずに表示できる非圧縮の TIFF に変換している。

PDF は、ファイルサイズが小さいため頒布しやすい。一方、非圧縮の TIFF はファイルサイズが大きいため、広く頒布するファイルとしては不便である。しかし、デジタル化資料のマスターデータとしては利用の目的や活用方法を限定しないファイルフォーマットが望ましい。

実際、配信用画像データは PDF をフォーマット変換して公開しており、今後、情報技術の発展によりさらに別の配信用画像フォーマットに変更して利用することも考えられる。デジタル化資料のマスターデータは時代や用途によって、どのデータフォーマットが最適かを判断し、適切なフォーマットに変換できる必要がある。

表 2　画像データの作成仕様

	初期			中期		
	2013	2014	2015	2016	2017	2018
画像データフォーマット	マルチPDF （Jpeg圧縮） 圧縮率の定めなし			TIFF（非圧縮）		
ファイル名	原資料の資料コード			原資料の資料コード-ページ番号（0000から始まる4桁の整数）		
トリミング	文書ではない背景はトリミングする。					
背景色	白色に色調補正			撮影時の背景のまま		
資料画像下部	画像の下部に余白を作り、原資料の資料コード-ページ番号（0000から始まる4桁の整数）　と、作成者名を付加。			なし		
メタタグ	-			作成者：沖縄県公文書館		
格納媒体	可搬ハードディスク					

2. 画像データの長期保存

作成した画像データは可搬ハードディスクに格納して引渡される。しかし、ハードディスクは通電せずに保管しておくと、数年でデータの読み取りができなくなる可能性があることや、媒体の寿命が 10 年程度であることから、長期保存には適さない。そのため、作成した画像データは、正本はブルーレイディスク、副本は磁気テープ（以下「LTO」という。)、2 式の長期保存用媒体に記録し、保存している。

正本のブルーレイディスクは､日本工業規格「電子化文書の長期保存方法（JIS Z 6017;2013)」（以下「本規格」という。）に準拠し、作成している（表 3)。

本規格に準拠する意義は、記録品質の高いディスクの作成ができ、記録後の初期検査により、ディスクの状態が数値により可視化され、記録後も長期的に見読性が確保された良好なディスクであることが担保されることである。また、定期検査では、利用による劣化及び経年劣化も数値化されるため、マイグレーションの時期を感覚に頼ることなく検討することができる。本規格については、宮長貴旨

の論考[1]に詳しい。本規格には、５年を目安に定期検査をすることが規定されているが、ディスクの利用頻度や先の検査値に応じて検査をする運用が望ましいと考える。

表 3　長期保存用ブルーレイディスクの作成仕様

媒体形態/品名	ブルーレイディスク　TL（100GB）
フォルダ構成	1媒体に複数の原資料の画像データを収録
ファイルシステム	UDF2.5以上
ディスク名	10桁の英数字（新規資料コード）
書込み速度	4倍速
ファイナライズ	する
記録後の検査	1ファイルごとのバイト数の照合 フォルダ名・ファイル数の確認 日本工業規格「電子化文書の長期保存方法（JIS Z 6017;2013)」に規定された初期検査結果が基準を満たしていること。

副本のLTOの作成仕様を表4に記す。LTOは、一度に大容量のデジタルデータを記録できるため光ディスクに比べて効率が良い。また論理的に改編可能であるため、過去に撮影を省略した頁を補う必要のある資料画像データの保存には便利である。反面、真正性を確保するために、作成時の照合に用いた1ファイル毎のバイト数は台帳等により保管し、画像データの更新に応じて記録をとる必要がある。

表 4　磁気テープの作成仕様

媒体形態/品名	LTO　Ulitorium7（6TB）
フォルダ構成	1媒体に複数の原資料の画像データを収録
ファイルシステム	LTFS
ディスク名	8桁の英数字（新規資料コード）
記録後の検査	1ファイルごとのバイト数の照合 フォルダ名・ファイル数の確認

3. 公開システム

公開システムは、主に「琉球政府文書デジタルアーカイブ」ウェブサイトのトップページや目録情報検索画面及び資料画像ビューア等のフロントエンドと、目録情報データベースや配信用画像データ等のバックエンドで構成されている。2016年（平成28）度の改修では、これらの公開システムをオンプレミス[2]に移行した。

1　宮長貴旨「長期保存用光ディスクの保存性能に関して」『日本写真学会誌77巻1号（社団法人　日本写真学会2014)』p 10-14

2　一般に、オンプレミスは運用主体がサーバを資産として持ち、サーバの専門知識を有する技術員を確保しなければならないなどの制限がある。しかし、サーバの設計から維持管理のすべてを運用主体で行えることから、必要なカスタマイズを自由に行うことができる利点を持つ。一方、クラウドサービスを利用したシステム運用は、初期投資が抑えられ、必要なストレージに応じた費用のみで運用できる上、運用主体がサーバという資産及びサーバの専門知識を有する技術員を持つことなく、ネットワーク環境さえあればどこでも運用することができる。公文書館は、2001年（平成13）より、沖縄県公文書館所蔵資料管理システムARCHAS21の目録情報を館内サーバにおいて運用し、ウェブ公開している。

3-1. 目録データベース

3-1-1. データベースの統合

改修前の公開システムは、すべてをクラウドシステムで構築しており、公文書館内サーバで既に運用をしている沖縄県公文書館所蔵資料管理システム ARCHAS21（以下「アーカス」という。）との連携がなかった（図 3）。そのため、改修前システムのウェブサイトを直接利用するユーザーは、琉球政府文書と関連の深い米国収集資料や沖縄関係資料、沖縄県文書について、別のウェブサイトで検索しなければならない。逆に、アーカスでは、本事業でデジタル化され、公開されている資料の特定ができない。つまり、初めからデジタルアーカイブを閲覧する意思のあるユーザーは、資料画像をウェブで閲覧できるが、公文書館ウェブサイト内の所蔵資料検索にて検索しただけでは、デジタル資料に辿り着くことができなかった。そこで、2016 年（平成 28）度の改修で、改修前システムの目録情報データベースをアーカスに統合した（図 4）。これにより、ユーザーは、公文書館所蔵資料検索からデジタル資料の閲覧ができるようになった。また、現在、整備途中であるが、資料画像ビューアの html に目録情報を持たせることで、ユーザーは公文書館のホームページを介さず、Yahoo! や Google 等のウェブ検索エンジンから資料画像を閲覧できる仕組みにしている。

目録情報データベースを統合した利点はユーザー利便性のみではない。公文書館はアーカスを既にオンプレミスな運用をしているため、目録データベースのクラウドサービスに掛かる運用コストを削減できた。

図 3　改修前の公開システム

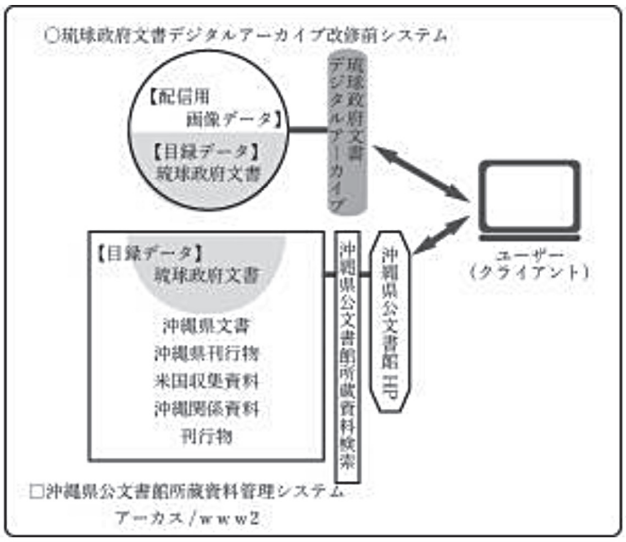
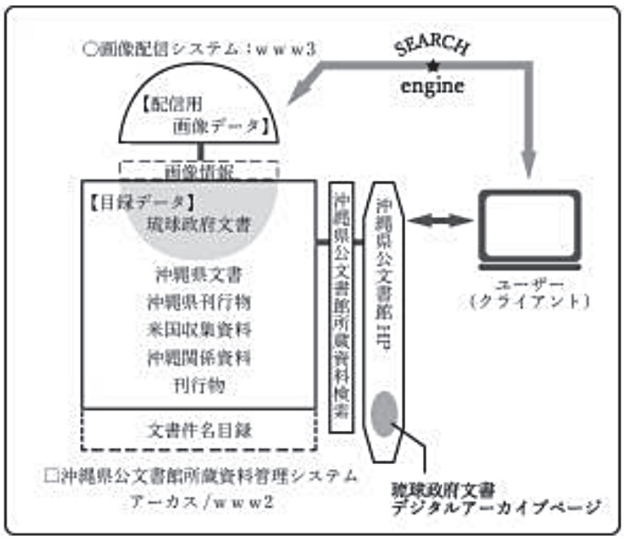


図 4　改修後の公開システム



3-1-2. 文書件名目録の追加

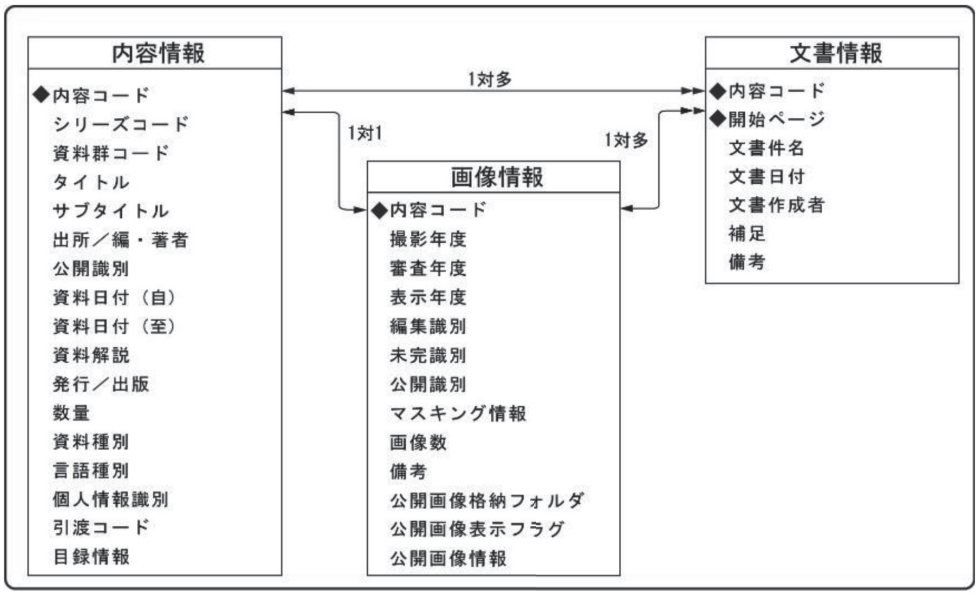
アーカスは、公文書館の所蔵資料を管理する複数のテーブルで構築されたリレーショナル型データベースを用いている[3]。それにより、多数の資料群の階層化が実現され、所蔵資料検索で目的の資料を特定しやすいシステムとなっている[4]。しかし、琉球政府文書の多くは、簿冊式ファイリングで整理・保管されており、１資料の中に複数の文書が存在しているため「雑書綴」といった抽象的な資料のタイトルの場合、その資料の内容が資料群の階層構造だけでは把握できず、アーカスで目的の資料を

3　大城博光「公文書目録情報のデータベースモデル～階層構造を持つ目録情報のリレーショナルデータベースでの実装～」『沖縄県公文書館研究紀要第 2 号』（沖縄県公文書館 2000）

4　豊見山和美「公文書目録データベースにおける階層構造の表現に関する試み～琉球政府文書を例に～『沖縄県公文書館研究紀要第 3 号』（沖縄県公文書館 2001）

特定できない。そこで、さらに検索精度を高めるため、文書の件名及び件名を補足するキーワード等で構成した「文書情報」テーブルを作成し、「内容情報」（公文書館所蔵資料の内容に関する情報）テーブルと関連付けた（図 5）。加えて、画像配信サーバとの連絡テーブルである「画像情報」と「文書情報」を関連付け、沖縄県公文書館所蔵資料検索で抽出した文書目録の該当画像をピンポイントで表示できるような仕組みにした。

図 5　追加したテーブルと既存内容情報との関連



3-2. 資料画像ビューア

改修前システムの資料画像ビューア「Opa View」（以下「Opa View」という。）は、左に資料目録データベースから取得する目録情報、右に資料画像を配置し、２つを連動していた（図 6）が、改修後の公開システム資料画像ビューア「Contents View FLEX」（以下「Contents View FLEX」という。）は、目録情報の詳細を公文書館所蔵資料検索結果詳細ページに任せ、資料画像ビューアには資料タイトルのみを表示している。追加機能として、Contents View FLEX の右には資料全体の視認性を高めるためのサムネイルを加えた（図 7）。

図 6　OPA View 画面設計書（2015）

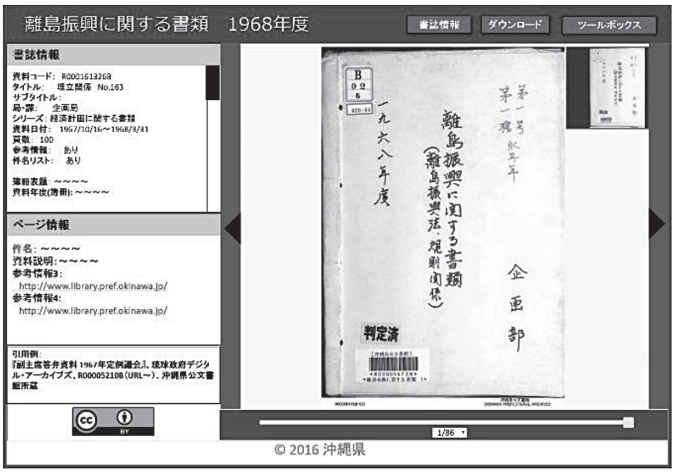
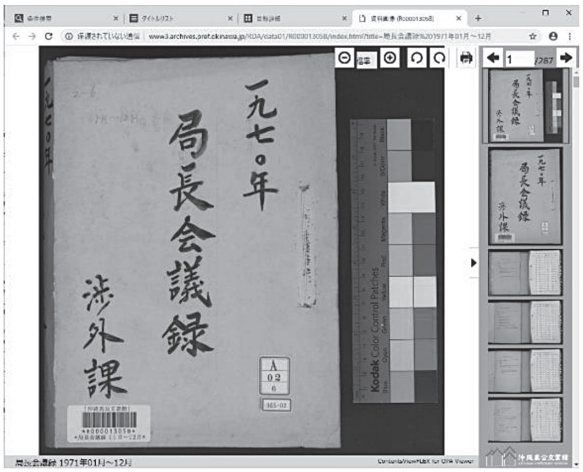


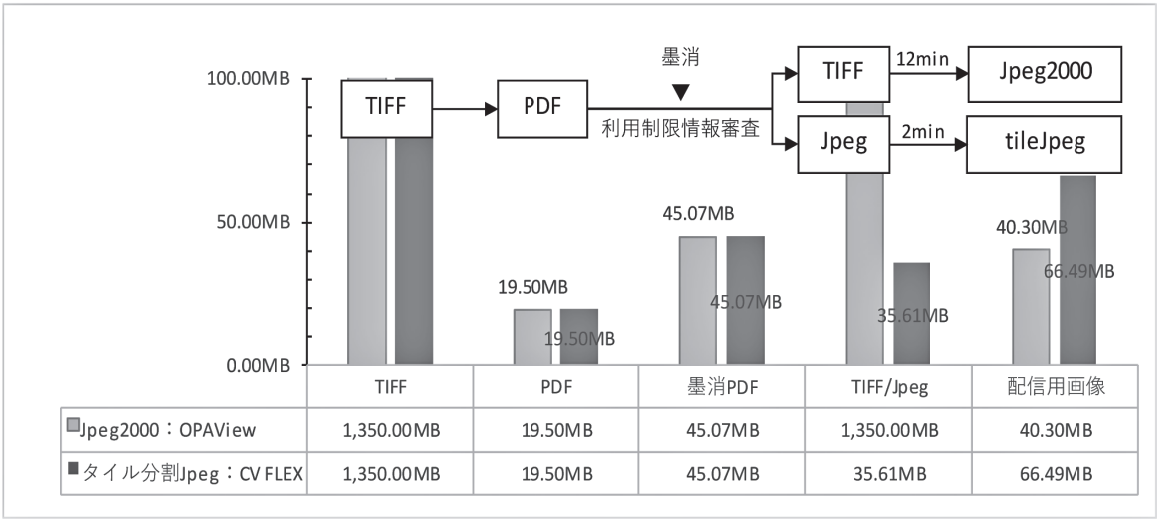
図 7　ContentsViewFLEX



画像表示部分は、Opa View 及び Contents View FLEX のいずれもブラウザ標準の言語である Javasprict で記述されたオープンソースの Open Seadoragon を利用しているが、公開システムの構築方法と配信用画像データフォーマットの違いから、ユーザー利便性と運用コストの両視点のパフォーマンスが異なる。

Opa View の配信用画像サーバは、クラウドに配信用画像 Jpeg2000-part1（以下「Jpeg2000」という。）をアップロードするだけで、ユーザーが使用する PC（以下「クライアント」という。）からのリクエストに応じて動的に Deep Zoom 形式のタイル画像を生成することができる。また、Jpeg2000 は Jpeg より圧縮による劣化を少なく、高度な圧縮技術でファイル容量を小さくすることができる。つまり、少ないストレージで高精細な画像を提供することができる（図 8）。反面、Opa View が採用している配信用画像サーバは、クライアントからのリクエストの都度、画像を生成するため、クライアントに届くまでに時間がかかっていた。Contents View FLEX の場合、あらかじめタイル分割 Jpeg（Deep Zoom 形式のタイル状に分割した Jpeg 画像）を作成し、画像配信サーバに用意する必要があるが、サーバはクライアントからのリクエストがあれば、リクエストの都度新たに画像を生成せずとも高画質な画像を提供できるため効率がよい。

図 8　画像データフォーマットの変換工程[5]とファイルサイズの推移



	TIFF	PDF	墨消PDF	TIFF/Jpeg	配信用画像
Jpeg2000：OPAView	1,350.00MB	19.50MB	45.07MB	1,350.00MB	40.30MB
タイル分割Jpeg：CV FLEX	1,350.00MB	19.50MB	45.07MB	35.61MB	66.49MB

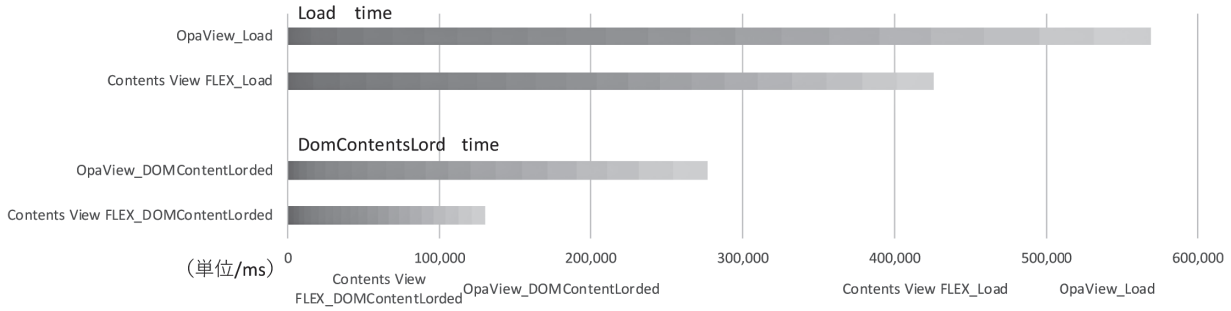
また、表示速度は配信用画像にフォーマットの違いだけで長短するのではなく、資料画像ビューアの構造も影響している。Opa View の場合、資料画像ビューアに目録情報を持たせているため、画像の表示に加え、ページを繰る度目録情報をデータベースへ取りに行く時間が発生していた。Contents View FLEX の場合は、サムネイルを持つため、検索ページから 1 ページ目を表示する際はサムネイル画像を load する時間が必要となる（2 ページ目以降はキャッシュをみるため必要としない）が、改修前に比べて少ない時間で資料画像の配信ができる（図 9）。なお、配信速度の比較には 2015 年（平成 27）度にインターネット公開した資料 125 冊のうち、一資料あたりのファイル数および、ファイルサイズが平均的な資料として『日本政府援助関係資料　昭和 45 年度　概算新規要求明細総括表』（R00004272B　沖縄県公文書館所蔵）を用いた。また、表示速度はユーザーのネット

5　資料画像データは、利用制限情報の審査のため、まず TIFF から作業用の PDF に変換し、資料に利用制限情報がある場合、その部分を墨消によってマスキングをする。その後、配信用画像に変換する工程を採っている。なお、比較に使用した画像は、普通紙、朱肉付普通紙、色入りの普通紙、大判文書、トレーシングペーパー、青焼等、異なる種類の文書画像を 20 枚抽出し、測定した。

ワーク環境により異なるため、測定には Google Chrome 開発者ツールを使用し、ネットワークの条件を “Regular3G(Download:750kb/s,upload:250kb/s,Latency:100m)” に設定した。

Opa View はクラウド運用のため、サーバ保管容量を抑える必要があることから、TIFF ではなく、Jpeg2000 に変換していた。また、Contents View FLEX は 1 枚の Jpeg でも表示可能であるが、大判文書等の物理的に大きな画像を表示する際、より高画質な画像を提供することを考え、画像検証をした上でタイル分割 Jpeg を採用することとなった。それぞれの配信用画像を選択する際の検討過程は省略する。

図 9　画像表示速度の比較

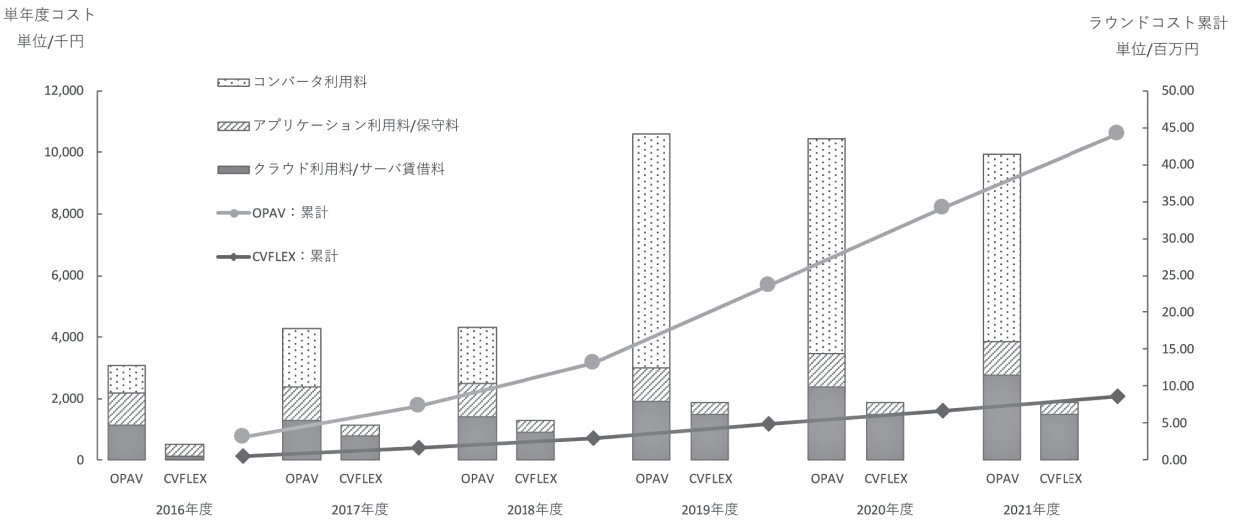


（単位/ms）	Contents View FLEX_DOMContentLorded	OpaView_DOMContentLorded	Contents View FLEX_Load	OpaView_Load
検索画面-1page	3,420	3,550	16,800	15,510
1~2page	3,698	4,010	17,147	16,750
2~3page	3,976	4,590	17,494	17,990
3~4page	4,254	5,740	17,841	19,230
4~5page	4,532	6,970	18,188	20,390
5~6page	4,810	8,110	18,535	21,610
6~7page	5,088	9,360	18,882	22,870
7~8page	5,366	10,530	19,229	24,060
8~9page	5,644	11,650	19,576	25,250
9~10page	5,922	12,740	19,923	26,480
10-100jump	6,200	13,695	20,270	27,386
100~101page	6,478	14,446	20,617	28,506
101~102page	6,756	15,365	20,964	29,390
102~103page	7,034	16,137	21,311	30,620
103~104page	7,312	17,013	21,658	31,493
104~105page	7,590	18,133	22,005	32,883
105~106page	7,868	19,105	22,352	33,882
106-200jump	8,146	19,966	22,699	34,982
200-250jump	8,424	21,046	23,046	35,918
250~251page	8,702	22,009	23,393	36,762
251~252page	8,980	23,029	23,740	37,683

3-3. 公開システム運用コスト

2018 年（平成 30）10 月末現在、本事業において公開している資料数は 13,500 点、2,051,232 ページであり、現運用公開システムにおける画像配信サーバのストレージ使用量は約 7.3TB である。これを OPA View を使って運用をした場合、約 4.38TB に抑えることができる。しかし、改修前システムを運用する場合のサーバストレージ使用料及びシステムを利用するためのアプリケーション利用料はコンバータ利用料（TIFF から Jpeg2000 にフォーマット変換するために必要な費用）を除いても､オンプレミスの場合のサーバストレージ賃借料及び運用保守費用の約 1.6 ～ 3 倍必要となる（図 10）。サーバストレージ及びアプリケーション利用料は琉球政府文書デジタルアーカイブシステムを使用する間の必要なランニングコストとなるため、避けて通れない指標である。

図 10　公開システム運用に掛かるラウンドコスト（概算）



公開（予定）資料数とストレージ使用量（概算）

	2015年度	2016年度	2017年度	2018年度	2019年度	2020年度	2021年度
公開資料数	125	2,903	6,251	6,000	25,000	23,000	20,000
資料枚数	19,022	441,776	951,273	913,076	3,804,481	3,500,123	3,043,585
公開資料数：累計	125	3,028	9,279	15,279	40,279	63,279	83,279
資料枚数：累計	19,022	460,798	1,412,071	2,325,147	6,129,628	9,629,751	12,673,336
ストレージ(CVF)	69.22GB	1,607.45GB	3,461.29GB	3,322.31GB	13,842.96GB	12,735.53GB	11,074.37GB
ストレージ(OPAV)	185.00GB	964.47GB	2,076.78GB	1,993.39GB	8,305.78GB	7,641.32GB	6,644.62GB

＊ストレージ容量及び費用は2018年10月末現在の数値から試算。

おわりに

本事業は開始当初から試行錯誤を経て進められている。デジタル化の目的そのものも、ウェブ上での公開画像の作成という限定的なものから、原資料の保存を考慮したマスターデータの作成へと変化し、マスターデータの長期保存も工程化されている。ウェブ上での公開システムも、ユーザー利便性を高め、かつ運用コストを低減する方向で改修された。

デジタル化による資料の完全な代替物の提供は、原本の利用を抑制して保存性を向上させる。また、デジタルアーカイブ化することによって、原資料のもつ情報を発信し広く活用される契機が生まれる。このように、本事業における現在のデジタルアーカイブは、保存と活用の両立を視野に入れている。

本事業は、沖縄県公文書館が所蔵する琉球政府文書を守り、普及する重要な事業である。琉球政府文書は、現在の沖縄の在り方を規定する過去の統治の実態を検証するために不可欠な資料であり、沖縄県公文書館はその永続的な活用を保障しなければならない。本稿がその過程で少しでも役立つことがあれば幸いである。