

にいがた「郷土新聞画像データベース」の構築 Development of Niigata Regional Newspaper Database

津田 光弘

Mitsuhiro Tsuda

イパレット, 大阪府豊中市清風荘 2-8-30

iPallet, 2-8-30 Seihusou, Toyonaka, Osaka

あらまし: 新潟県立図書館の「郷土新聞画像データベース」は、明治から昭和初期に新潟で発行された地方新聞の閲覧提供を目的とし、県立図書館内とその承認機関で 2017 年の公開を予定している。現時点で発行単位で約 3 万件、画像数 20 万点超を年月日、紙名、発行区分などで検索し閲覧できる。本データベース構築の概要と、特に IIIF Image API と専用画像ビューアを用いた紙面の表示方法、表示範囲の自動記録機能をつけた記事見出しのテキスト編集を紹介し、最後に新聞データベース共通の課題を考察する。

Summary: The purpose of "Niigata Regional Newspaper Database" of Niigata Prefectural Library is to provide regional newspapers issued in Niigata from the Meiji period to the early Showa period. It is scheduled to be released this from the library and its approval agency at 2017. At present, the database manages about 30,000 issue units, exceeds 200,000 images, and can be searched and browsed according to issue date, newspaper name, issue type, etc. We introduce an overview of this development, especially, an image viewing method using the IIIF Image API with a dedicated image viewer, and a text editing of the article headline with an automatic area record function. Finally, we discuss common problems in the newspaper database.

キーワード: 新聞、画像、地域資料、データベース、IIIF Image API

Keywords: newspaper, image, regional materials, database, IIIF Image API

1. はじめに

新聞は日々の出来事を広範囲に記録した速報性に優れた資料であると共に、長期的な社会状況や人々の暮らしの変化、広告などからもその時代の雰囲気を知ることができる特徴をもつ資料分野である。とりわけ、各都道府県において発行される地方新聞は、特定の地域について調べる際に重要である。

新潟県立図書館(以降、県立図書館と記載)では、所蔵する新潟日報(旧新潟新聞)、毎日新聞(旧東京日日新聞)など 11 タイトルの地方新聞の資料提供を、従来マイクロフィルムで行っていたが、そのデジタル化を行い、明治から昭和初期に新潟で発行された一部を「郷土新聞画像データベース」として構築した。現時点で 3 紙、発行単位で約 3 万件、紙面数で 20 万超の画像を、年月日、新

聞紙名、発行区分などで検索し閲覧できる。県立図書館内と、その承認機関で 2017 年の公開を予定している。

新聞や雑誌などの刊行物の公開では、画像数が大量になり、そのデータ配信や管理の問題に直面する。また、マイクロフィルムを経由してのデジタル化という点でも充分な解像度が得られ難く、旧字を含む活字印刷による画像でもあり、紙面内容のテキスト化は困難といえる。必然的に検索は発行日などの情報に頼らざるを得ないが、とはいえ、計画的かつ効率的な検索用テキストの蓄積や整備も望まれる。

本稿では、この「郷土新聞画像データベース」の構築の概要と、特に大量の画像の配信に用いた技術、記事見出しテキストの編集機能を紹介し、最後に新聞データベース共通の課題を考察する。

2. 背景

2.1. 「にいがた MALUI 連携・地域データベース」

今回報告する「郷土新聞画像データベース」は、新潟県立図書館と新潟大学人文社会・教育科学系附置地域映像アーカイブ研究センターの「にいがた MALUI 連携・地域データベース(Niigata MALUI Regional Database)」^[1]に属する。このデータベースの現在の構造と運営組織を図1に示す。

ここで、MALUI とは、従来の MLA(博物館・図書館・文書館)に加え大学・産業の連携を示すが、これらの機関が所蔵する文化資源をデジタル化し、一元的に管理公開するシステムによって、研究における資料活用の効率性・利便性を高めることの意義は、今日改めて説明の必要はないであろう。加えて、地域に特化した統合データベースは、新潟の歴史・社会の情報を豊富に含み、地域への貢献としても期待される。

なお、地域映像アーカイブ研究センターでは、新潟県内各所から発掘された映像や音声資料を、図1における「にいがた 地域映像アーカイブ・データベース」として従来から公開してきた^[1]が、今回そのリニューアルに合わせ、MALUI 連携に取り組んだ。

2.2. 「郷土新聞画像データベース」について

新潟県立図書館では、『新潟新聞』をはじめとした明治時代以降の郷土新聞を所蔵しており、従来マイクロフィルムの形で資料提供を行ってきた。新潟日報(旧新潟新聞)、毎日新聞(旧東京日日新聞)など11タイトルの所蔵マイクロフィルムを閲覧でき、複写申請を行える。

その所蔵する新聞マイクロフィルムのうち、明治から昭



図1. データベースと組織

和17年までに発行された新潟県関係の新聞の一部をデジタル化し、検索・閲覧できるサービスを「郷土新聞画像データベース」として公開する予定である。

これにより、資料の閲覧が容易になるとともに、新潟県立図書館の館内だけでなく、新潟大学附属図書館や県内市町村立図書館等のデータベース利用の承認を受けた施設で、順次 Web 経由での閲覧が可能となる。

郷土新聞画像データベースの公開予定資料とその数量を表1に示す。

表1. 公開対象となる新聞とその数

(2017年1月現在)

新聞名	記事の年月日	発行日件数※	紙面数※
新潟新聞	1877(明治10)年4月～1941(昭和16)年7月	17,718 +	108,758 +
新潟毎日新聞	1917(大正6)年1月～1941(昭和16)年7月	8,479	73,024
新潟時事新聞	1924(大正13)年2月～1933(昭和8)年3月	3,292	24,118
新潟日日新聞	1941(昭和16)年8月～1942(昭和17)年10月	予定	予定

※ 件数、紙面数は、データベースからの検索値で、破片や不特定紙面を含む。

2.3. 新聞画像データについて

紙面の原画像データは、県立図書館において次の工程で製作された。

- (1) マイクロフィルムから専用スキャナで自動読込
- (2) スキャナ付属のアプリケーションを用い、自動接続
- (3) 紙面単位で JPEG ファイルとして保存

画像は8bit グレースケールの中品質の JPEG であり、概ね文字の判読可能な解像度範囲にある。画像サイズは、時代によって比率の違いがあるが、長辺の画素値は4千～9千、データ量は平均で6 MByte(4～12 MByte)である。画像の品質としては、階調のばらつきはあるが概ね良好な範囲、周辺はマイクロ撮影時に焦点が合っていないものもある。

ファイル名の表記構造は表2のようで、その紙面の基本的情報がファイル名自体から得られる。他に、ファイル名と、新聞名、発行年月日が記載された一覧ファイルがある。データベースの製作時には、原画像との対応関係を維持するため、このファイル名を公開用画像データの識別名として原則そのまま用いた。

表2. ファイル名の構造

例: a19260405a07

a	1926	04	05	a	07
紙名	西暦年	月	日	発行種別	面数

コード	紙名	紙名補足
a	新潟新聞	
b	新潟日報	新潟社
c	日刊新潟	
d	新潟時報	
e	新潟毎日新聞	
f	新潟時事新聞	
g	新潟日日新聞	
h	新潟日報	新潟日報社
i	タリニイガタ	
j	タリ新潟日報	
k	新潟日報タリ	

コード	発行種別
x	不定(※2)
a	朝刊
b	夕刊
c	附録
d	号外
n	午後版
t	追加
i	判別不可能
k	広告(※2)

※1 不明箇所はアンダースコアを用いる
※2 製作者が追加

3. システムの全体構成

本画像データベースの全体構成と製作の推移を図2と表3に基づき説明する。

開発工程は概ね2期に分かれる。1次構築ではデータの把握とシステムの内部構造の設計を行った。試験システムを作成し検証を行った上で、配信方法や画像データのフォーマットを確定した。

予定の内覧までには開発期間があったが、それよりも以前にプロジェクトのサーバーの新規導入に合わせて配信技術の確定をする必要があった。配信用の画像をサーバーに登録依頼する必要もあった。画像データの総量が3TB とかなり大きいため、ネットワーク経由での配置が実際上不可能であったことに依る。

2次構築では主に公開や運用に向けたユーザインタフェースの開発とパフォーマンスの検討を行なった。

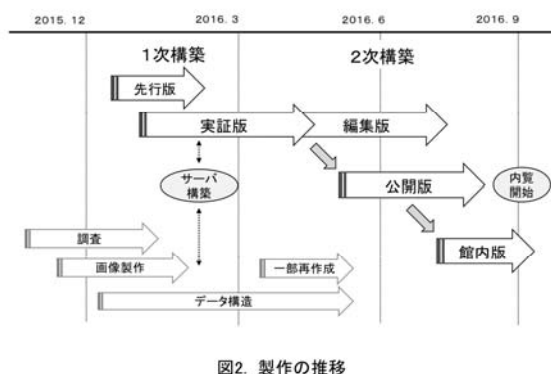


図2. 製作の推移

3.1. 1次構築

まず、画像の品質チェックや時代毎の新聞の特徴などを把握する目的で、公開対象となるすべての紙面をざっと閲覧できるWeb環境を製作した。

画像データのリストをプログラムで取得し、提供されたファイル名の一覧と整合チェックを行なった。一部に不一致なども見つかったが、それらは後工程の課題として保留し、データベースの前段階となるテーブルを作り、それを用いて簡単なWebページを作成した。このバージョンを先行版と名付けた。

次に、先行版の一覧テーブルからリレーショナルデータベースを作成し、実証用の動的サイトを構成した。

この実証版の目的は、

- ・データベースの基本的な構築
- ・検索方法や応答速度の確認
- ・画像サーバー(後述)の動作試験
- ・画像配信プロトコル(同)の検証
- ・新聞用画像ビューア(同)の動作確認
- ・各種機能の事前検証

等であり、2次構築の基礎となる。実証版で検証した各項目や機能については、本稿の後半で詳説する。

3.2. 2次構築

本画像データベース・システムは、図3のシステム概要で示すようにデータを管理する部分は共通であるが、公開方法や記事編集、管理の面を考慮して、3種類のユーザインタフェースを用意した。このうち、公開版と館内版は利用者向けで、編集版はデータベースの管理や記事の見出し編集を目的とした非公開のバージョンである。

表3. 各バージョンの内容

バージョン	目的	画像配信	備考
先行版	全データの確認	静的HTML JPEG	目的終了
実証版・編集版	画像配信方式の確認 記事見出し編集	IIIF Image API 画像ピラミッドTIFF	利用者認証
公開版	内覧用 —— 共通デザイン	分割画像	解像度制限
館内版	正式公開用	IIIF Image API 画像ピラミッドTIFF	

3.3. 公開版

内覧公開用に製作したバージョンで、幅広い利用者による操作検証を目的とする。新聞の発行情報に基づき、検索と表示だけが行える。サイトの外観や機能は MALUI 連携データベースとして共通なデザインとしている。「にいがた 地域映像アーカイブ・データベース」とは実装に用いたプログラム言語は異なるが、共通のスタイルシートを適用し、操作手順なども可能な限り近づけた。

画像表示は見出しテキストが判読できる程度の低解像度としている。共通かつ試験的な機能として、HTML5 のローカルストレージを用いた個人用の紙面選択・登録機能と検索履歴機能を用意した。

図 4 に公開版の画面遷移と操作の手順を示す。検索の内容も図中に表示する。

3.4. 館内版

公開版とほぼ同一のユーザインタフェースであるが、県立図書館の館内端末での利用を主に想定して設計した。紙面の選択・登録機能と検索履歴機能は、不特定多数が操作する設置環境上、無効化している。詳細画像の閲覧とその印刷機能を用意した。

また、将来的な仕組みであるが、記事見出しのテキスト検索を行える。

3.5. 編集版

1次構築の実証版のユーザインタフェースを、ほぼそのまま拡張した。特に、後述する記事見出し編集の用途に特化して機能を追加した。

この編集版によって、共通に利用するデータベースを管理しデータの編集を行う。ユーザー認証、ユーザー管理機能を備える。編集版の検索・表示インタフェースについては図 6 に示す。検索とその結果の行き来がしやすいように単一ページで操作が完結する。利用はタブレットでも行える。

4. 画像の配信と閲覧技術

4.1. 画像配信方法

紙面の原画像データは解像度としては判読の下限付近であるが、新聞というメディアの性質上、文字が細かく、紙面が大きいので、画像としては書籍や雑誌などの刊行物に比べ画素数はかなり大きい。

Web で配信するには今日のネットワーク環境であっても

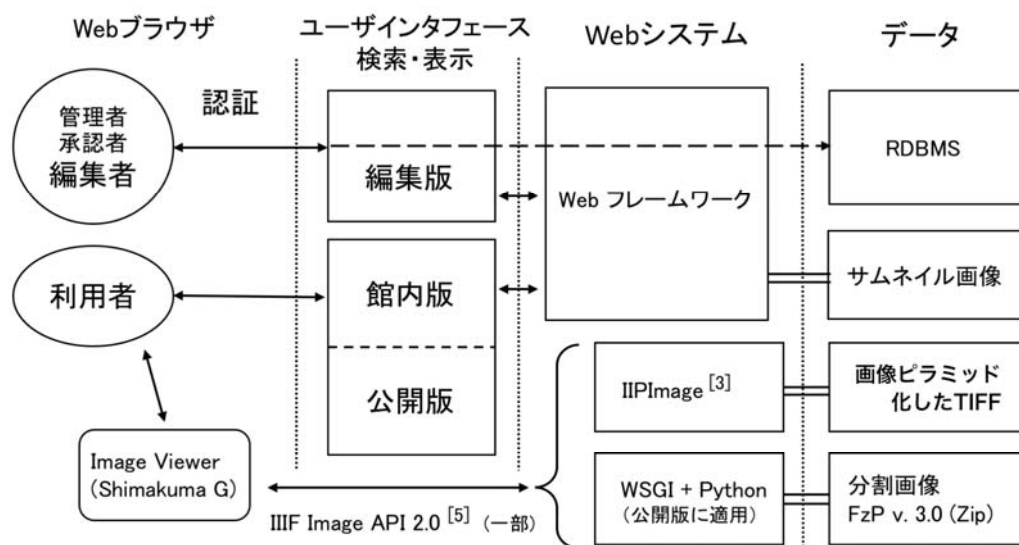


図3. 郷土新聞画像データベースのシステム概要



説明：四角数字：主な表示・操作の流れ、丸数字：機能や操作ボタン

図4. 公開版のユーザインタフェース

単板の JPEG や PDF としては表示までの待ち時間もかかるため、著者としては多階層多分割の分割画像 (FreezoomPack, Deep Zoom, Zoomify など) の形式にし、サーバーに大量の小さな JPEG 画像を配置する擬似的なストリーム配信手法を候補とする場合が多い。しかし、原画像総数が製作時点で 20 万以上であり、これを変換すると 1 億レベルのファイル数になる。しかも、今後が増えるため、サーバー上での管理や将来的なデータの移設なども考慮し調査を行った。仮想環境によるシミュレーション検証を重ね、配信手段には画像ピラミッド化した TIFF (Tiled Pyramidal TIFF) と、対応するオープンソースの画像サーバー (IIPImage, ライセンスは GPL) を採用した。^[3]

利点としては、

- ・大量の分割されたファイルを配置しなくて良い
- ・TIFF 内部では JPEG 圧縮するため、遠隔でも後からの画像の追加や差し替えが比較的容易
- ・画像サーバーを経由するため Web 側から直接に画像ファイルをダウンロードできない

一方、課題は、

- ・プロジェクトのサーバーに関連ライブラリを含めたインストールが必要で、OS によっても構築法が異なる
- ・アクセス集中時のサーバー負荷
- ・配信用画像の制作方法
- ・将来的な持続性

画像ピラミッド化した TIFF は、オープンソースの画像ライブラリ libvips (ライセンスは LGPL 2.1+) ^[4] を用い、Mac OS X 10.11 上でバッチ製作した。制作後の総データサイズは、JPEG 圧縮した状態で約 1.3TB である。

なお、サーバー負担を想定し、補助用に分割画像の FreezoomPack でも製作した。その他、一覧とプレビュー用に2種類のサムネイルも制作した。これらの画像はすべて libvips とそのプログラム拡張を実施して行った。

4.2. 画像閲覧の方法

画像配信サーバーとして用いた IIPImage は、多階層多分割の配信を行えるサーバーである。1990 年代に提案された Internet Imaging Protocol で URL により画像の一部

あるいは全体を取得できる他、近年、画像の国際的な相互利用を推進するための枠組みとして提唱されている IIIF (International Image Interoperability Framework) Image API にも対応する。^[5]

IIIF コンソーシアムはこの目的に沿った規格や画像サーバー、画像ビューアも案内している。この仕組みを用いると、画像フォーマットの種類にかかわらず必要な画像を図 5 のような URL で要求できる。このプロトコルに対応するように画像サーバーや画像ビューアを実装することで、フォーマットやサーバーへのアクセス方法の違いに煩わされることなく画像を利用できる。技術や製品の多様性を維持しつつ、既にアーカイブ化された画像資産を公開持続にする今日的な標準化だと考えることもできる。

この将来的な可能性を検証するため、Image API 2.0 (現在は 2.1) の一部を用いて定型のタイル画像を呼び出す専用の画像ビューアを開発した。

この画像ビューアはこの他、FreezoomPack (v3.0 以降) や Zoomify、Deep Zoom といった分割画像フォーマットにも対応するマルチプロトコルな画像ビューアである。

最大の利点は IIIF Image API を用いることで、これらの分割画像へのアクセスが画像ピラミッド化した TIFF を用いた場合も含めて共通化でき、特定の画像フォーマットの制約から解放されることである。公開版では Zip 圧縮状態の FreezoomPack (v3.0) から WSGI (Web Server Gateway Interface) 経由で画像を取得する仕組みも配信のバックアップシステムとして平行して構築し、安定性評価を行っている。

```
{scheme}://{server}/{prefix}/{identifier}/  
[region]/[size]/[rotation]/[quality].[format]
```

例: FreezoomPackフォーマットの、sample画像に関して
範囲(0,0,1024,1024)を256画素の正方タイルとしてJPEGで読み込む

```
http://ipallet.org/opossum/sample.fzp/  
0,0,1024,1024/256,256/0/default.jpg
```

図5. IIIF Image API 2.0 画像リクエスト URI シンタックス ^[5]

5. 記事見出し編集ツール

明治から昭和初期にかけての新聞をテキスト検索できれば活用範囲は広がるが、マイクロフィルムを経た画像でもあり、現時点では識別率とコストの両面で自動的なテキスト化は現実的ではない。とはいえ、テキスト化を進めなければ、記事の集合体である大量の画像から有効な箇所を見つけることは難しいままである。

そこで、新聞の記事見出しを登録・編集する試験的なツールを編集版に追加した。通常のテキスト入力に加え、記事の範囲を取り込む仕組みを設けた。図 7 に編集画面を示す。範囲を指定するには、一般にマウスなどでドラッグ操作で定義する方法を思うが、今回の方法では、対象とする記事を表示し、その見出しをキーボードで入力し登録する際に、表示している範囲(四隅の座標)を自動的に見出しテキストの関連情報として記録する。

再現時にはこの座標データと紙面画像の画素数を用いることで、拡大率を計算し画面中央に表示できる。

記事見出し編集時には、編集者が表示範囲を意識する必要があるが、これによって編集時のみならず閲覧時の利用者負担が軽減できることを意図した。

また、記事見出しの編集では、利用者に編集者と承認者の権限を設け、簡単な承認レベルを設けている。上位権限である承認者のチェックを経て公開が可能になる。さらに上位の管理者が管理システムで各権限を利用者に割り当てる。

記事見出し編集の製作や確認においては、県立図書館内で実施されていたデータ約数千万件を用いた。

編集ではこのほか、承認権限以上において、紙面自体の情報の編集や記事の分類設定なども行える。これらの編集用インタフェースは完成ではなく、今後の計画の中で随時調整してゆく予定である。

6. 考察

6.1. 新聞の公開方法について

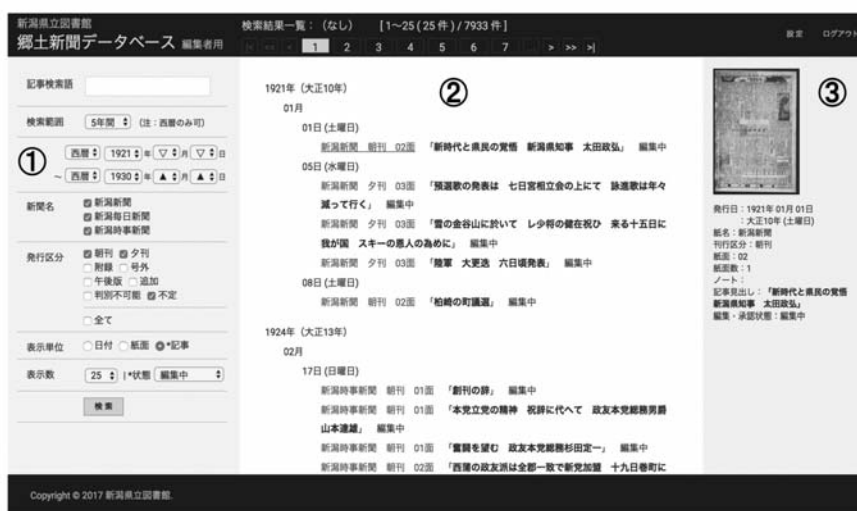
過去の新聞は、その資料が成立した時代の社会状況や雰囲気を知ることのできる重要な歴史的資料である反面、現在においては適切でないと思われる表現や出来事等

も記されている。また、現在でも個人のプライバシーに関わる内容もありうる。このため、一般に新聞資料は館内でのみ公開というポリシーが設定されることが多い。

一方、研究調査において、資料群の存在は公開されているが、その内容の目録などが公開されておらず現地へゆくまで内容がわからないという指摘もある。^[2]

新聞の場合、構造として記事の見出しがある。テキスト

検索はまだできないが、見出しだけでも画像として確認ができれば目録の代わりになる。その上で、より詳細に調べたい場合は現地まで足を運んでいただければ良い。そのような考え方で、公開版では、記事本文の判読までは拡大しても難しいが、記事見出しは読める解像度レベルの画像の公開を提案した。



① 検索, ② 一覧, ③ プレビュー

図6. 編集版の検索・一覧画面



①: 取得範囲 (= 表示範囲全体), ②: 記事範囲 ①: 一覧と新規作成, ②: 編集ツール

図7. 編集版の紙面ビューアと記事編集

6.2. 大量の画像のハンドリング

一度に約20万点の、しかも、ある程度のサイズを有する画像を扱うことは著者にとっては初めてのことであったが、その製作で直面した問題についてまとめる。

- ・全貌の把握が極めて困難。

特に、画像データの破損、ファイル名などの例外の発見が困難で画像変換、並べ替えなどの作業をして、初めて問題が見つかることも多い。

- ・負荷のため動作しないテキストエディタがある
- ・配信用画像の製作、転送に時間がかかる

画像製作に相当の日数がかかり、サムネイルなどの解像度変換や画像コピーにもかなりの時間がかかる。対応できるプログラムを製作しながらこれらの作業を行なったが、途中でシステムトラブルも何度か経験した。コンピュータパワーは一昔前に比べ格段に進歩したが、製作環境の調整は怠ることはできない。

6.3. 大きな画像の配信フォーマット

今回は、紙面の一覧のしやすさを目的に画像ピラミッド化したTIFFを用いたが、PDFという選択肢は資料の配信目的、特に印刷を意図する場合にはある。著者の経験上も、本報はこのフォーマットでの新聞資料の公開を排除する報告ではないことは追記しておく。

6.4. 新聞データの画質について

著者は従来から何度か新聞画像を扱うことがあったが、今回の画像は解像度、階調ともに比較的に良好な状態であった。画像欠損等のトラブルもわずかにあったが、手作業で修復できる範囲であった。製作された県立図書館の担当者の苦勞を思う。

これまで他組織の話として感じていたことで、稀に、マイクログラフからのデジタル化の際に、どのような理由によるものかモノクロ画像で製作がされる場合がある。しかし、自動的な2値閾値の決定は難しく、解像度にも限界があるので判読が極めて困難になる。このような2値化は、記録メディアも安価になった今日デジタル化で選択するべきではないと考える。もし、検討中の場合はご考慮願いたい。

7. まとめ

新潟県立図書館の「郷土新聞画像データベース」の構築を通じ、閲覧目的や公開制限の違いに対応するシステムと、3種類のユーザインタフェースを実現した。

紙面数20万超という大規模な画像配信を実現するため、IIIF Image API 対応の画像サーバーと、その規格の一部に対応する専用画像ビューアを用いて、低負荷な状態であるが安定的な配信を検証できた。

また、この画像ビューアは新聞の記事見出し編集機能を備え、テキスト編集に加え、結果的にではあるが、記事のおおよその範囲を自動記録する。これによって、記事検索後に紙面上の目的の位置に直ぐに到達できる仕組みを実現した。これは新聞というメディアに適した方法であることを確認した。画像データベースは今後、公開に向けて必要な調整を行なってゆく予定である。

謝辞

本データベースの製作は、新潟大学人文社会・教育学系と新潟県立図書館と著者との実証研究として行なった。配信サーバーにリモート接続できる許可をいただき、普段は決断しにくい画像サーバーや配信方法などの手法を検討できた。この機会と各面でご支援をいただいた皆様に感謝したい。

参考文献

- [1] 新潟大学人文学部. 「こいがた 地域映像アーカイブ」 No. 5, ISSN 1883-5643, 2014.
- [2] 金城厚. 民謡研究と伝承のための録音資料アーカイブの展望. 「人文科学とコンピュータシンポジウム」論文集 2015. pp. 199-202.

参考サイト

- [3] IIPImage: <http://iipimage.sourceforge.net/>
- [4] libvips: <https://github.com/jcupitt/libvips/>
- [5] IIIF: <http://iiif.io/>