2019年3月11日発行

津波ディジタルライブラリィにおける新聞記事検索システムの改良および海外からのアクセス対応の検討

今　井　さやか

相模女子大学紀要　　VOL.82（2018年度）

津波ディジタルライブラリィにおける新聞記事検索システムの改良および海外からのアクセス対応の検討

今　井　さやか

Improvement of Tsunami Digital Library for query of newspaper and access from foreign countries

Sayaka IMAI

Abstract

We are developing a Tsunami Digital Library (TDL) from 2003. In TDL, we provided digital books, newspaper contents, videos, CG simulations of Tsunami disaster etc. by using PC web browser. Since TDL has been running 15 years, we have three problems in operation. The first is aging of a TDL system. The second is that query of newspaper articles are not enough for users. The third is corresponding to access from foreign countries. In this paper, we focus on the second and the third problems. We reconsidered key words for Tsunami disaster, and improve database query system in TDL. Also by use of Google Website Translator, we correspond to access from foreign countries.

Key Words：Tsunami Digital Library, Newspaper Articles, Website Translator - Google Translate.

１．はじめに

我々は津波に関する文献、新聞記事、津波遡上シミュレーション動画、フィールドワークデータ、津波災害ビデオなどを総合的に管理し、インターネットを通じて広く公開する津波ディジタルライブラリ（Tsunami Digital Library: TDL）を開発し、その活用についても提案を行っている[1][2][3]。2003年から現在までつづくプロジェクトであり、常にデータの追加が行われ、継続的に運営されている。筆者らは2011年の東日本大震災後も、津波災害記録だけでなく、復興計画の策定の際に、過去の情報に遡ることのできる資料のディジタル化を行い、ライブラリィの継続的稼働に努めてきた。近年では、2016年熊本地震、2018年北海道胆振東部地震および台風による沿岸地域の高潮被害など、自然災害が相次いで発生しており、TDLは膨大な過去の記録を参照できる存在として重要な役割を今後も期待されると思われる。

TDLが長期にわたって継続的に稼働するうえで、現在問題となってきたことが３点ある。１点目はTDLを稼働するWebサーバの老朽化などシステムの問題、２点目は資料の検索方法、３点目は海外からのアクセス対応である。このうち、１点目のWebサーバの老朽化については、東北大学災害科学国際研究所のアーカイブプロジェクト「みちのく震録伝」[4]の協力を得て、プロジェクトのサーバシステムおよびネットワークリソースを利用してTDLの稼働を続けることが可能となった。2018年度にTDLシステムの移行作業が完了し、今後も継続してTDLを稼働する環境を整えることができた。２点目の資料の検索方法については、特に、明治から昭和時代にかけての津波災害に関する新聞記事の検索方法に対して、記事のディジタル化の際に付与し

たキーワードや津波災害名のばらつきにより、検索要求に十分な結果が得られていない点である。改めて、津波災害名を整理し、新聞記事のキーワード付けを見直し、津波災害ごと、新聞社ごとに整理して、日付、紙面のページ番号で整理してアクセスできるようにした。３点目の海外からのアクセスに対応つ

いては、まず文献リストのWebページにGoogle翻訳ツール[5]を利用し、文献のタイトルを英訳して表示できるよう改修を行った。

以降、本稿では２章で新聞記事の検索システムの改修について、３章では、TDLのサイトでのGoogle翻訳ツールの利用について報告する。

図１　改修後のTDLトップページ

２．新聞記事検索システムの改修

2．１　地震津波災害名の整理

図１は改修後の津波ディジタルライブラリィの検索トップページである。このページの左のカラム「文献一覧」から各種津波災害に関する文献にアクセスすることが可能である。製本された本の形式で発行された文献および、地震津波被害に関する新聞記事を文献の種類に応じて「論文」、「報告書」、「雑文」、「新聞記事」としておおまかな種類に分け、それぞれの文献リストにアクセスすることが可能である。論文、報告書、雑文のリンクからはそれぞれのカテゴリの文献のタイトル一覧を閲覧することが可能で、文献のタイトルリンクから目的の資料のテキストデータおよび図表画像にたどることができる。一方、新聞記事のリンクからは、津波災害ごとに記事を整理し、それぞれの津波災害の新聞記事をたどれるようになっている。すべての文献と新聞記事に地震津波災害名がキーワードとして付与されているが、その名称については、文献や新聞記事で多く使用されている名称を採用しており、文献の数や新聞記事が多くなるにつれて、付与された名称のばらつきが目立つようになってきた。ディジタル化された際の津波名で分類されているため、例えばチリ地震津波に対する新聞記事は「チリ地震」「チリ地震津波」「チリ地震地震（キーワード付与時のミス）」と

いう異なる表現が存在していた。そこで、津波災害の名称を改めて検討し、新聞記事の整理を行った。

TDLの中での統一した津波災害の名称については以下のようにした。

- 1896年明治三陸沖地震
- 1933年昭和三陸沖地震
- 1944年東南海地震
- 1946年南海地震
- 1960年チリ地震津波

津波名の名づけルールについては、発生西暦年＋地震名とし、地震名は理科年表[6]の名称を参考にした。ただし、1896年（明治29年）および1933年（昭和８年）と２回発生している「三陸沖地震」に対しては、理科年表では「明治」「昭和」という文言は名称にはついていないものの、「明治三陸沖地震」「昭和三陸沖地震」などと様々な文献やメディアで使用されていることからそれぞれ明治、昭和の単語を付与した。このようにルール付けすることで、上記以外の津波災害も統一した表現で整理することが可能となる。

2. 2　新聞記事のテキストデータ化

TDLに掲載されている新聞記事のリンクからは、津波被害ごとに、全国紙と災害が起きた地域の地方紙の複数の新聞社の記事を閲覧することができる。新聞記事は津波災害が発生した日から約１か月間の津波災害に関する記事をテキストデータ化している。新聞記事は、新聞の１面をディジタル化の最小単位とし、その面に掲載されている地震津波災害関係記事のみテキストデータ化し、記事タイトル、本文、新聞社名、日付、面番号などの情報のXMLタグを付与している。以下は、新聞記事に付与しているXMLタグの例である。

- メタデータ（新聞記事そのものの情報）
  - title：文献（新聞）名
  - creator：発行新聞社名
  - subject：文献のキーワード（津波災害名、新聞の日付の元号表記・西暦表記）
  - publisher：発行者
  - date：新聞の発行年月日
  - type：文献のジャンル（Newspaper）
  - format：text
  - identifier：文献ID（西暦日付８桁の数字＋２桁の面番号）
  - language：言語
  - rights：文献に適用される権利に関する情報
- 新聞記事本文
  - num：面番号
  - article：記事
    - title：記事タイトル
    - text：記事テキスト

TDLには、新聞１面を最小単位として1691件の新聞記事テキストデータを掲載している。図２に1896年明治三陸沖地震についての、東奥日報の1896（明治29）年６月17日３面の記事の画像（地震津波に関する記事部分を切り抜いた画像）を示す。また、図３に図２の画像から記事をテキスト入力し、XMLタグを付与したXMLデータを示す。TDLのデータベースにはこのXMLテキストデータを格納する。

図２　地震津波災害に関する新聞記事
（記事部分のみを切り出した画像）

```
<?xml version="1.0" encoding="UTF-8"?>
- <newspaper>
  - <metadata>
      <title>東奥日報</title>
      <creator>東奥日報社</creator>
      <subject>明治三陸地震津波</subject>
      <subject>明治29年6月15日</subject>
      <subject>1896年6月15日</subject>
      <publisher>東奥日報社</publisher>
      <date>1896-06-17</date>
      <type>Newspaper</type>
      <format>text</format>
      <identifire>1896061703</identifire>
      <language>ja</language>
      <rights>津波ディジタルライブラリィ"http://tsunami-dl.jp/"</rights>
    </metadata>
  - <page>
      <num>03</num>
    - <article>
        <title>　　●湊村の海嘯 </title>
        <text> 一昨夜来数十回の地震、何処にか■災のなからん やと、人の噂もとりどりなりしが、昨朝八戸より電 報によれば、昨夜八時頃（十五日）海しゅう震い来たり、三 戸郡湊村大字湊字白銀の民家、流失するもの四戸、 破壊するもの八戸、学校一棟、その他死亡者三名、 生死知れざるもの二名、船舶漁具等の流失破損 数多なりしと。右報に接するや、鈴木警部長は松本 課長を従え、昨日の二番列車にて実況視察のため 同地に出張せりと言う。なお上北郡海岸にも被害 ありし由にて、目下取り調べ中なりと。 </text>
      </article>
    - <article>
        <title>　　●一昨夜来の地震 </title>
        <text> 当地にて此の一両日■■震あり、一昨夜七時三十三 分三十秒より五分間弱震あり、やや強かりしため 人々戸外に避けたるもありき。それより引き続き八 時十分に一分間の微震あり。後三回の極微震につい で、八時五十九分二十秒より二分間の微震あり。その 後六回の極微震の後、九時五十六分三十秒よりニ 分間の微震は、南北に掛けたる時計の運転止まる 程なりしが、その後六回の極微震あり■昨日午前零 時四十八分四十五秒より三分間の微震につき、四 回の極微震、午前四時十七分には三分間の微震、三 回の極微震、同八時四十九秒より四分間弱震あり、 同九時四十六分十二秒に微震、続いて極微震三回。 都合弱震二回、微震六回にして、右は昨日午前十一 時頃の調べなるが、何れも南西北東に地平■なり き。 </text>
      </article>
    </page>
</newspaper>
```

図３　XMLタグ付けされた新聞記事

2．3　新聞記事検索結果一覧表示の作成

新聞記事のデータベース検索については、トップページに設置したGoogle検索を利用したフリーワードによる検索の他、<subject>タグに記述されている津波災害名による検索を準備した。問題となっていた津波名のばらつきは、システムであらかじめ対応表を作成し、検索の際に読み替えることで津波災害ごとの検索を実装した。新聞記事は以下の順にソートされるようにした。

- 北海道新聞社
- 東奥日報社
- 奥羽日日新聞社
- 岩手公報社
- 岩手日報社
- 河北新報社
- 東京朝日新聞社
- 毎日新聞社
- 横浜毎日新聞社
- 中日新聞社
- 伊勢新聞社
- 徳島新聞社
- 紀州民報社

上記の新聞社の順に記事を並べ替えて、さらに同じ新聞社の中で、<identifier>タグに記載している文献ID（西暦日付８桁の数字＋２桁の面番号）において並べ替えて、一覧表示する。図４に「1896年明治三陸沖地震」をキーワードに検索し、前述の新聞社の順、<identifier>タグに記載している文献IDの順にソートした結果一覧を示す。図５は図４の一覧から選択した新聞記事の表示例である。図３のXMLテキストデータをデータベースから検索し、加工して表示している。元のキーワードは「明治三陸地震津波」となっているが、統一したキーワード「1896年明治三陸沖地震」で検索されることが確認できる。

図４　TDLにおける新聞記事の閲覧ページ

図５　新聞記事ページ

３．海外からのTDLへのアクセス対応

3．1　TDLへの海外からのアクセス対応の現状

TDLに掲載されている文献には英語で記述された文献もあるものの、トップページをはじめ多くのWebページが日本語のページとなっている。また、図６に示すように、英語で記述されたWebページ[7]を公開し、動画の解説文の英訳や、津波防災として重要な文献の英訳を掲載している。しかし、英訳の手間が負担となり、コンテンツの公開は一部にとどまっている。このように、海外からのアクセスに対しては十分な対応ができていないことが課題となっている。近年、インドネシア・スマトラ島の地震津波災害をはじめ、日本だけでなく海外でも津波が発生し、その後の復興や災害対策において、過去の記録や論文を保持するTDLは役に立つと思われる。

現状のTDLに設置しているGoogle アナリティクス[8] を用いたアクセスログの分析においては、90%以上のアクセスユーザが日本語を使用するユーザであり、海外（英語圏他）のユーザは５％程度にとどまっている。

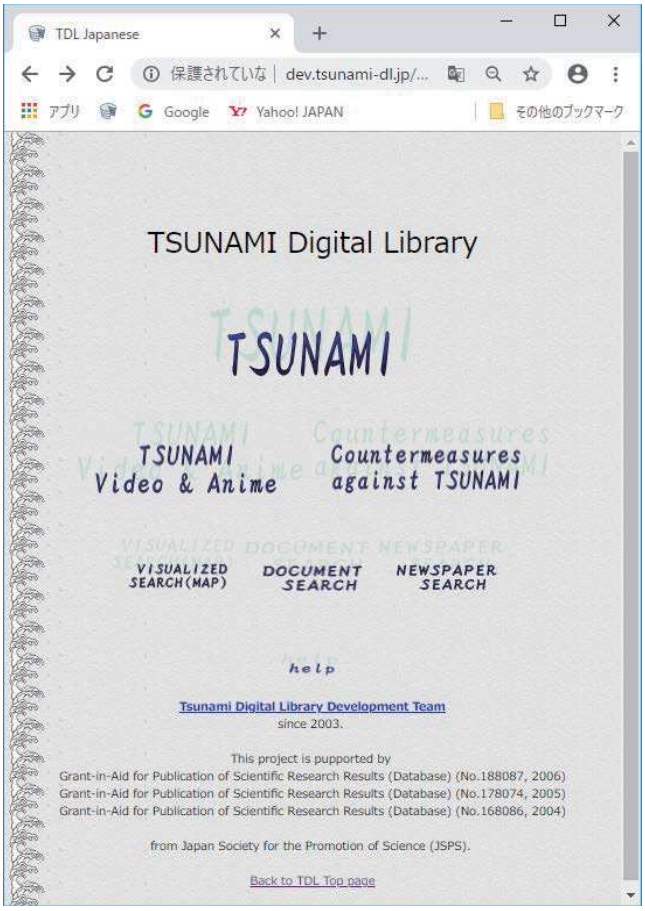

図６　TDLの英語サイト

3. 2　Googleウェブサイト翻訳サービスの利用

本研究では、海外からのアクセスユーザに対する対応として、まずは、Googleウェブサイト翻訳[4] を実行するツールボタンを主なWebページに設置した。図１に示した改修後のWebページの左上に設置されているツールボタンが翻訳を実行するボタンであり、現時点ではTDLのウェブサイトで日本語／英語を切り替えることができる。図７に言語の切り替えの様子を示す。翻訳ツールボタンを設置するには、Googleのウェブサイト翻訳ツールのページにおいて、Googleアカウントを使用して、翻訳するウェブサイトのURL、ウェブサイトの元の言語を設定し、プラグインの設定、翻訳する言語の選択などを行い、最終的にコードが自動生成される。発行されたコードを、Webページのツールボタン設置場所に挿入することで、翻訳ボタンの設置が完了する。図８にTDL用の翻訳ツールボタン設置用のコードを示す。TDLでは、トップページ、およびトップページにリンクが掲載されている「論文」、「報告書」、「雑文」、「全文献（論文・報告書・雑文）」の各文献リストのページに設置し、英訳されたページを表示できるように実装した。図９にトップページのGoogleウェブサイト翻訳による英語訳ページを示す。

図７　Google翻訳ボタンから「英語」を選択

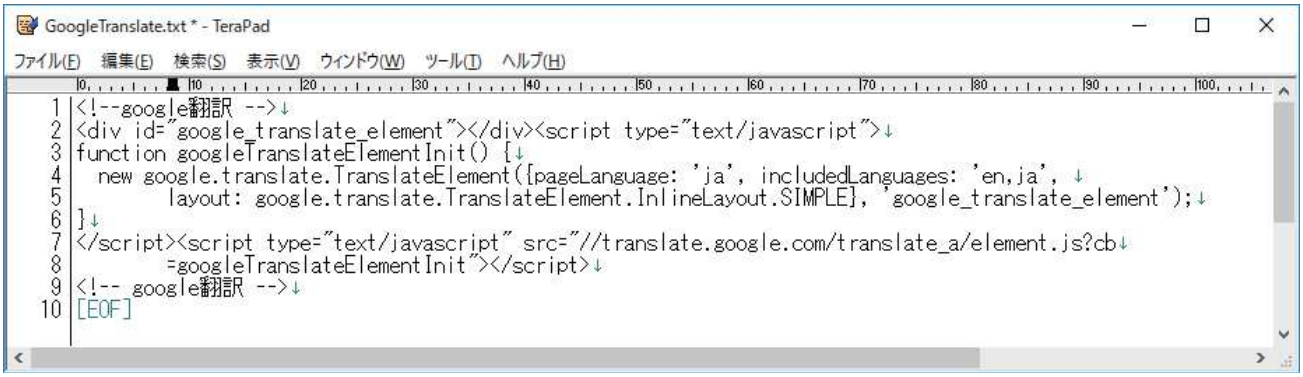

図 ８　Googleウェブサイト翻訳ツールによるHTMLコード

図９　Google翻訳によるトップページの英訳

４．まとめ

本研究では、長期にわたり稼働していることで、津波ディジタルライブラリィの課題となっていた新聞検索の改良および、海外からのアクセスに対応するためのWebサイトの翻訳ツールの設置について報告した。今後は、引き続き、文献や新聞記事の検索機能の強化を検討し、また、翻訳ツールの設置により海外からのアクセスの動向がどう変化するか注目したい。

謝辞

本研究は、平成30年度東北大学災害科学国際研究所リソースを活用した共同研究助成、および平成30年度相模女子大学特定研究助成費(A)を受けて実施されたものである。

参考文献

[1] Tsunami Digital Library: http://tsunami-dl.jp
[2] Sayaka Imai, Yoshinari Kanamori and Nobuo Shuto: Tsunami Digital Library, J. Gonzalo at al. (Eds.) ECDL2006, LNCS 4172, pp.555-558, 2006. Springer-Verlag Berlin Heidelberg 2006.
[3] Sayaka Imai, Yoshinari Kanamori and Nobuo Shuto: A Public Education Tool for Tsunami Disasters Based on Walking Tours in TDL, Proceedings of the 2010 JCDL, pp.377, 2010.
[4] 東北大学アーカイブプロジェクト　みちのく震録伝：http://shinrokuden.irides.tohoku.ac.jp/
[5] ウェブサイト翻訳ツールGoogle翻訳: https://translate.google.com/manager/website/?hl=ja
[6] 国立天文台編：理科年表 平成30年 第91冊，丸善．
[7] 津波ディジタルライブラリィ英語サイト：http://tsunami-dl.jp/old-content/TSUNAMI/TDL_top_e.html
[8] Google Analytics: https://developers.google.com/analytics/?hl=ja