Vol.2017-NL-233 No.7
2017/10/24

全地方議会会議録の横断検索に向けたデータ収集とデータ構造の検討

井原大将[1,a)]　内田ゆず[2,b)]　高丸圭一[3,c)]　木村泰知[4,d)]　江崎浩[1,e)]

概要：全国には都道府県・市・特別区・町・村を合わせて，1788 の地方自治体が存在しており，このうち約 86%がウェブ上で地方議会会議録を公開している．しかしながら，ウェブ上での会議録の公開方法やデータ形式は自治体により異なっており，横断検索や集計などにおいてそれらを統一的に扱うのは難しい．そこで本稿では，収集や整理がしづらい会議録を対象として，それらのデータ収集方法における問題点とその対処法について述べ，横断検索や集計が可能となるデータ構造を提案する．

Hiromasa Ihara[1,a)]　Yuzu Uchida[2,b)]　Keiichi Takamaru[3,c)]　Yasutomo Kimura[4,d)]
Hiroshi Esaki[1,e)]

1. はじめに

近年，地方自治体による公共データのウェブ公開，いわゆるオープンデータが進められている．地方政治における議論の過程がすべて記録された地方議会会議録も活用が期待される公共データの一つである．地方議会会議録の活用を推進する取り組みでは，複数の自治体の会議録を横断的に取り扱うことが要求される．たとえば「地方議会議事録横断検索 chiholog」では，798 の自治体から収集した地方議会会議録の全文検索や単語出現頻度の時系列分析機能等をウェブサービスとして提供している．この取り組みはオープンデータの活用例であるとして，VLED 利活用普及委員会勝手表彰において表彰を受けている[*1]．また，木村らは地方議会会議録コーパスを用いた学際的応用研究を進めており，情報工学，社会言語学，計量経済学などの分野で成果をあげつつある [1][2][3][4][5]．このような地方議会会議録を対象としたウェブサービスやコーパスを構築するためには，地方自治体がウェブサイトにおいて個別に公開している会議録を収集し，共通のデータ形式に整理した上で，データベースに登録する必要がある．しかしながら，会議録のウェブ公開方法は統一されていないため，公開方法に応じた収集・整理の手法を用意する必要がある．地方議会会議録のウェブ公開は，全文検索システムによる公開と，PDF 等のファイルによる公開に大別される．上述のウェブサービスやコーパスは現段階では，前者のみを対象としている．本研究の最終的な目標は，多様な形式で公開される地方議会会議録をすべて収集し，横断的な分析を可能にすることである．このため，PDF 等で公開される会議録も収集・整理の対象とする必要がある．そこで本稿では，まず地方議会会議録の公開状況と公開形式について述べる．次に，さまざまな形式で公開される会議録を統一的に扱うためのデータ構造の提案と，実際のデータ収集において生じた問題への対処について述べる．最後に，提案したデータ構造を用いて行ったデータ収集・整理の結果について述べる．

[1] 東京大学
The University of Tokyo
[2] 北海学園大学
Hokkai Gakuen University
[3] 宇都宮共和大学
Utsunomiya Kyowa University
[4] 小樽商科大学
Otaru University of Commerce
[a)] taisyo@hongo.wide.ad.jp
[b)] yuzu@eli.hokkai-s-u.ac.jp
[c)] takamaru@kyowa-u.ac.jp
[d)] kimura@res.otaru-uc.ac.jp
[e)] hiroshi@wide.ad.jp
[*1] https://bitlet.co/

2. 地方議会会議録の Web 公開の状況と公開形式

地方議会会議録の公開には大きく分けて，会議録システムとファイルの 2 つの形式がある．

2.1 会議録システムによる公開

会議録システムとは，議会の会議録を Web 上で閲覧/検索できるようにした Web システムであり，このシステムを通して Web ページとして会議録が提供される（図 1 に千葉市議会の例を示す）．ユーザは開催日や議会の種別，発言者，発言内容などを指定して検索し，閲覧する．会議録システムは，基本的には地方議会ごとに異なる，ソフトウェアベンダーによって開発されたシステムを導入している場合が多い．主要な会議録システムを表 1 に挙げる．

表 1 主要な会議録システム

会議録システム	ソフトウェアベンダー
DiscussNet	NTT アドバンステクノロジ
DB-Search	大和速記情報センター
Sophia	神戸綜合速記
VOICES	フューチャーイン

図 1 会議録検索システムによる公開の画面例（千葉市議会）

2.2 ファイルによる公開

ファイルによる公開では，地方議会や地方自治体のウェブサイト上で，ファイルとして会議録が提供される（図 2 に茂原市議会の例を示す）．ユーザはウェブサイトにアクセスし，ファイルをダウンロードして閲覧する．多くの場合，ファイルの形式は PDF ファイルや Microsoft Word ファイルである．

図 2 ファイルによる公開の画面例（茂原市議会）

3. Web 公開の会議録のデータ収集における問題

Web で公開されている会議録を収集するには，解決しなければならない 2 つの問題がある．これらの問題はどちらの公開形式にも共通である．

3.1 会議録のリソースをいかに見つけるか

Web 公開において，会議録のデータは地方議会のウェブサイト内のファイル（PDF ファイルや Microsoft Word ファイル）や会議録システム上の Web ページとして提供される．しかしながら，それらはユーザが Web ブラウザを通してアクセスすることが前提となっており，ユーザは視覚的なボタン配置やリンクテキストの情報に基づいて会議録のリソースへ辿りつく．そのため，コンピュータでクロールする場合は，クローラに会議録へのリソースへたどり着くための情報，例えば，リンクの辿り方や書き換え方などを与える必要がある．

3.2 会議録のデータをいかにデータ構造に落とし込むか

収集した会議録データは統一的なデータ構造で管理する必要がある．Web 公開において，会議録のデータのファイル形式，フォーマットは地方議会によって異なる．そのため，統一的に会議録を扱うためには，地方議会ごとの会議録のデータのファイル形式やフォーマットなどデータの構成を考慮し，データ構造を定義する必要がある．また，コンピュータでクロールする場合は，ファイル形式やフォー

マットの差異を吸収し，同じデータ構造にするための情報，例えば，”HTML のあるタグでマークアップされたテキストは会議名に変換する”等の変換ルールを与える必要がある．

4. 地方議会会議録横断検索の要件とデータ構造の定義

ここで，地方議会会議録横断検索の要件を下記に挙げる．

(1) 対応する地方議会すべてに対して同時に横断的に検索できる．
(2) 会議録の本文や会議名などを対象としたキーワード検索ができる．
(3) 検索条件として開催日や自治体名を使用できる．

次に，この要件のためのデータ構造を定義し，その定義を表 2 に示す．斎藤ら [6] はさらに細かい粒度のデータ構造を定義している．例えば，会議録本文を発言者，発言内容のフィールドに分割し，会議名の他に回や号などのフィールドを定義している．本研究では，すべての地方議会に対応するためのクローラの作成における実装上の難しさや，すべての地方議会に対応できるデータ構造の抽象化の難しさを考慮し，簡素なデータ構造を採用する．

表 2 地方議会会議録横断検索のデータ構造定義

項目	型	備考
地方議会 ID	数値型	地方議会を表す番号． 地方公共団体コードを用いる．
会議録本文	文字列型	
会議録会議名	文字列型	
会議開催日	日付型	

5. データ収集における問題への対処

会議録の収集では，地方議会の会議録の Web 公開ページにクローラでアクセスし，会議録データの探索と保存を行う．地方議会ごとに異なる URL や HTML マークアップをを用いているため，クローラは地方議会ごとに作らなければならない．しかし一方で，同じソフトウェアベンダーによる会議録システムや同じファイル形式を使っている自治体間では，URL や HTML マークアップに共通部分が多いこともある．そこで，今回は会議録システムごとのクローラとファイルによる公開に分けてクローラを作成する．本章では，クローラの具体的な動作を会議録の公開形式別に述べる．

5.1 会議録システムによる公開の場合

5.1.1 会議録データの探索

会議録システムは，会議録の検索と閲覧の機能を提供する．そして，会議録の閲覧や検索結果の表示のための URL はシステムによって動的に生成される．そのため，会議録の検索結果を表すページや会議録データを表すページのリンクには規則性がある．例えば，図 3 の例では，議事録データページの URL は `Template=DocOneFrame` というパラメーターを持っている．これを利用し，通常ユーザーが，検索機能によって目的の会議録一覧を表示し，そこから目的の会議録を閲覧する場合と同じように，クローラも同様の手順でアクセスを行う．具体的には，すべての会議録に対して検索条件を指定せずに検索を行い，すべての会議録の検索結果を表示し，個々の会議録データページへのリンクを取得する．ページが分かれている場合は，再帰的に次のページの会議録一覧を表示し，会議録リンクを取得する．会議録システムや地方議会によっては，会議録リンクが，会議録全文表示に対するリンクでなく，発言単位表示のリンクや HTML フレームへのリンクとなっている．その場合は，URL のパラメータを置換し，会議録本文表示への会議録リンクへと修正する．例えば，発言単位表示のテンプレートを指定するパラメータを会議録全文表示へ置換する．

```
 1 <TR>
 2   <TD align="right" nowrap><SPAN class="BasicRed">1</SPAN></TD>
 3   <TD><SPAN class="Basic">2017.04.17</SPAN></TD>
 4   <TD nowrap>
 5     <SPAN class="Basic">
 6       <A href="/chiba-c/dsweb.cgi/documentframe!1!guest01
         !!4744!1!1!1,-1,1!4629!312126!1,-1,1!4629!312126!2,0,2!4629!312126!319064!!1?
         Template=DocOneFrame" target="_self">平成
         29年新庁舎整備調査特別委員会 名簿</A>
 7     </SPAN>
 8     <SPAN class="Basic">
 9       <A href="?Template=DocOneFrame" target="_self"></A>
10       <BR>
11   </TD>
12 </TR>
13 <TR>
14   <TD align="right" nowrap><SPAN class="BasicRed">2</SPAN></TD>
15   <TD><SPAN class="Basic">2017.04.17</SPAN></TD>
16   <TD nowrap>
17     <SPAN class="Basic">
18       <A href="/chiba-c/dsweb.cgi/documentframe!1!guest01
         !!4744!1!1!1,-1,1!4629!312126!1,-1,1!4629!312126!2,0,2!4629!312126!319066!!2?
         Template=DocOneFrame" target="_self">平成
         29年新庁舎整備調査特別委員会　本文</A>
19     </SPAN>
20     <SPAN class="Basic">
21       <A href="?Template=DocOneFrame" target="_self"></A>
22       <BR>
23   </TD>
24 </TR>
25 <TR>
26   <TD align="right" nowrap><SPAN class="BasicRed">3</SPAN></TD>
27   <TD><SPAN class="Basic">2017.03.15</SPAN></TD>
28   <TD nowrap>
29     <SPAN class="Basic">
30       <A href="/chiba-c/dsweb.cgi/documentframe!1!guest01
         !!4744!1!1!1,-1,1!4629!312126!1,-1,1!4629!312126!2,0,2!4629!312126!326714!!3?
         Template=DocOneFrame" target="_self">平成
         29年第１回定例会（第９日目）議事日程・名簿</A>
31     </SPAN>
32     <SPAN class="Basic">
33       <A href="?Template=DocOneFrame" target="_self"></A>
34       <BR>
35   </TD>
36 </TR>
```

図 3 会議録検索システムの検索結果における URL の記述例（千葉市議会）

5.1.2 会議名抽出

検索結果のページにある会議録データページへのリンクテキスト（`<a>`タグのテキスト）や議事録データページのページタイトル（`<title>`タグのテキスト）を会議名として抽出する．

5.1.3 会議録本文抽出

会議録データページの HTML テキストから HTML タグを取り除き，本文を抽出する．ただし，一部の HTML タグが議事録の書式の一部となっている場合は必要に応じてその書式へ置換する．例えば，改行を表すタグ（`<br>`タグ）が該当する．

5.1.4 開催日抽出

会議名や会議録本文から抽出するが，開催日の表記には様々なパターンがあるため，それを考慮した抽出のルール（正規表現）を作成する．考慮すべき揺れや書式の例を次に示す．

- 半角数字や全角数字の混在
- 空白文字の混在
- 西暦表記，元号表記
- /（スラッシュ），-（ハイフン），，（カンマ）などの年月日区切り文字

5.1.5 JavaScript で動的に生成されるページコンテンツ

一部の会議録システムでは，会議録データを見つけるのに必要な情報が，JavaScript によって生成されることがある．例えば，図 4 の例では，議事録データページへの URL が JavaScript の `document.write()` によって生成されている．通常クローラは HTTP[S] クライアントとして動作し，サーバーに対して HTML ドキュメントを要求し，取得するだけであるため，Web ブラウザー上で生成されるコンテンツには対応していない．JavaScript で動的にコンテンツを生成する会議録システムや会議録システムのあるページに対しては，クローラ標準の HTTP[S] クライアントではなく，Web ブラウザーを外部からプログラムで操作して JavaScript を実行し，コンテンツを得る必要がある．

```
<SCRIPT type="text/javascript">
  <!--
    if( "1" == "1" ){
      document.write('<SPAN class="BoldRed">1<\/SPAN>');
    }else{
      document.write('<A href="/kisarazu-c/dsweb.cgi/query!1!guest02
      !!2200!1!0!1,-1,1!2986!209134!1,-1,1!2986!209134!2,0,2!2986!209134!!!1?Template
      =List&QuerySelect=No&List=now&DocumentType=" target="_self">1<\/A>');
    }
  //-->
</SCRIPT>
<SCRIPT type="text/javascript">
  <!--
    if( "2" == "1" ){
      document.write('<SPAN class="BoldRed">2<\/SPAN>');
    }else{
      document.write('<A href="/kisarazu-c/dsweb.cgi/query!1!guest02
      !!2200!1!0!1,-1,1!2986!209134!1,-1,1!2986!209134!2,0,2!2986!209134!!!13?
      Template=List&QuerySelect=No&List=now&DocumentType=" target="_self">2<\/A>');
    }
  //-->
</SCRIPT>
<SCRIPT type="text/javascript">
  <!--
    if( "3" == "1" ){
      document.write('<SPAN class="BoldRed">3<\/SPAN>');
    }else{
      document.write('<A href="/kisarazu-c/dsweb.cgi/query!1!guest02
      !!2200!1!0!1,-1,1!2986!209134!1,-1,1!2986!209134!2,0,2!2986!209134!!!25?
      Template=List&QuerySelect=No&List=now&DocumentType=" target="_self">3<\/A>');
    }
  //-->
</SCRIPT>
```

図 4　会議録検索システムの JavaScript によって生成されたコンテンツの記述例（木更津市議会）

5.2 ファイルによる公開の場合

5.2.1 会議録データの探索

ファイルによる公開では，地方議会の会議録公開 Web ページに一覧として会議録データへのリンクがあるか，階層的になっていて例えば，"平成 29 年会議録" → "定例会" → "第 1 回定例会第 4 号会議録本文.pdf"のようにリンクが階層化されていることが多い（図 5 に一宮町議会の例を示す）．そこで，クローラには地方議会ごとにリンク深さ，リンク URL フォーマット，リンクテキストフォーマットを与え，指定されたリンク深さのリンクをすべて走査し，そのうち URL のフォーマットとリンクテキストに合致するリンクを会議録データへのリンクとして取得する．URL のフォーマット（正規表現）の例としては".*\.pdf"，リンクテキストには"会議録"や"本文"などを指定する．

図 5　ファイルによる公開のリンクを辿る画面例（一宮町議会）

5.2.2 会議名抽出

先程述べた指定されたリンクの深さまで走査するときに，各ページにおいてページタイトル (<title>タグのテキスト）とリンクテキスト（<a>タグのテキスト）と直上の Heading テキスト (<h6>, <h5>, <h4>, <h3>, <h2>, <h1> タグのテキスト）を収集し，それらから，"定例会"や"回"，"号"を含むものを抽出する．例えば，図 5 の一宮町議会では各ページにおいて，ページタイトルが "議会会議録 | 一宮町役場" → "平成 28 年　会議録 | 一宮町役場"，リンクテキストが "平成 28 年　会議録" → "平成 28 年第 4 回定例会会議録"，Heading テキストが "議会会議録" → "平成 28 年　会議録" になり，会議名として "平成 28 年第 4 回定例会会議録" が抽出される．

5.2.3 会議録本文抽出

ファイル形式に応じて，変換ツールを用いて本文抽出を

行う．今回のクローラでは，PDF 形式のファイルのために PDFMiner[*2]を用いた．

5.2.4 開催日抽出

会議録システムの場合と同様の開催日抽出を行う．

5.2.5 ファイルのメタ情報を用いた情報抽出

ファイル形式によっては，メタ情報に会議名や開催日に結びつく情報が格納されていることがある．具体的には，PDF ファイルの場合は，タイトルやアウトラインに会議録の情報があることが多い．そのため，5.2.2 で述べた会議名抽出が失敗したときに代わりとして利用できる．

6. データ収集結果

先に述べたデータ構造とデータ収集に基いて，千葉県内の市町村議会の会議録を収集した結果を表 3 に示す．千葉県には地方議会が全部で 54 議会（37 市 16 町 1 村）あり，会議録公開に関する内訳としては，そのうち 49 議会（同県内全議会 91%）が会議録の Web 公開を行っている．公開形式でみると，会議録システムが 35 議会（同県内全議会 65%），ファイルによる公開が 15 議会（同県内全議会 28%）である．ファイルによる公開をしている議会のファイル形式は，すべて PDF ファイルである．作成したクローラによって，会議録システムによる公開をしている地方議会に関しては，市川市議会を除く 34 議会，ファイルによる公開をしている地方議会に関しては，勝浦市議会を除く 14 議会の会議録を収集できた．今回作成したクローラは大手 4 会議録システムに対応したもののみである．市川市は独自の会議録システムを採用しているため，会議録データの収集ができなかった（図 6 に市川市議会の会議録公開画面を示す）．勝浦市は 5.2.2 で述べた会議名抽出に失敗したため、会議録データの収集ができなかった．具体的には，表組みされたセルの中に，会議録へのリンクや会議名に関する情報が格納されており，ページタイトル，リンクテキスト，Heading テキスト内に会議名に関する情報が含まれていなかった（図 7 に勝浦市の会議録公開画面を示す）．この形式に対応するためには，表組みを考慮した会議名抽出を行う必要がある．旭市議会と九十九里町議会，長柄町議会，大滝町議会は 5.2.5 で述べたファイルのメタ情報として PDF のアウトラインに会議名が入っていたが，5.2.2 の会議名抽出で会議名を抽出できたため利用しなかった．同様に茂原市議会は PDF のタイトルに会議名が入っていたが，5.2.2 の会議名抽出で会議名を抽出できたため利用しなかった．

7. まとめ

本稿ではまず，地方議会会議録の Web 公開の 2 つの形式，会議録システムによる公開とファイルによる公開の形

図 6 市川市の会議録公開画面

図 7 勝浦市の会議録公開画面

について説明し，それぞれの公開形式に共通する収集における 2 つの問題を示した．そして，それらを解決するためのデータ構造の定義とデータ収集法について述べた．提案手法を千葉県内の地方議会に対して適用した結果，会議録システムによる公開をしている地方議会に関しては，35 議会中 34 議会，ファイルによる公開をしている地方議会に関しては，15 議会中 14 議会の議事録を収集することができた．今後の課題としては，提案手法の他の地方議会への適用が挙げられる．

[*2] https://euske.github.io/pdfminer/

表 3 千葉県内の市町村議会における会議録の公開形式とデータ収集結果

議会	公開形式	備考	収集結果
千葉市	会議録システム	DBsearch	o
木更津市	会議録システム	DBsearch	o
南房総市	会議録システム	DBsearch	o
山武市	会議録システム	DBsearch	o
長生村	会議録システム	DBsearch	o
白井市	会議録システム	DBsearch	o
銚子市	会議録システム	Discuss	o
館山市	会議録システム	Discuss	o
松戸市	会議録システム	Discuss	o
野田市	会議録システム	Discuss	o
成田市	会議録システム	Discuss	o
東金市	会議録システム	Discuss	o
習志野市	会議録システム	Discuss	o
柏市	会議録システム	Discuss	o
市原市	会議録システム	Discuss	o
流山市	会議録システム	Discuss	o
八千代市	会議録システム	Discuss	o
我孫子市	会議録システム	Discuss	o
鎌ケ谷市	会議録システム	Discuss	o
君津市	会議録システム	Discuss	o
浦安市	会議録システム	Discuss	o
四街道市	会議録システム	Discuss	o
袖ケ浦市	会議録システム	Discuss	o
印西市	会議録システム	Discuss	o
富里市	会議録システム	Discuss	o
匝瑳市	会議録システム	Discuss	o
香取市	会議録システム	Discuss	o
いすみ市	会議録システム	Discuss	o
大網白里市	会議録システム	Discuss	o
栄町	会議録システム	Sophia	o
船橋市	会議録システム	Voices	o
佐倉市	会議録システム	Voices	o
鴨川市	会議録システム	Voices	o
富津市	会議録システム	Voices	o
市川市	会議録システム	独自	x
茂原市	ファイル	PDF	o
旭市	ファイル	PDF	o
勝浦市	ファイル	PDF	x
八街市	ファイル	PDF	o
東庄町	ファイル	PDF	o
九十九里町	ファイル	PDF	o
横芝光町	ファイル	PDF	o
一宮町	ファイル	PDF	o
睦沢町	ファイル	PDF	o
長柄町	ファイル	PDF	o
長南町	ファイル	PDF	o
大多喜町	ファイル	PDF	o
御宿町	ファイル	PDF	o
鋸南町	ファイル	PDF	o
酒々井町	ファイル	PDF	o
神崎町	なし		x
多古町	なし	議会だよりあり	x
芝山町	なし	議会だよりあり	x
白子町	なし	議会だよりあり	x

参考文献

[1] 木村泰知, 関根聡: 主辞に基づく政治問題抽出手法, 人工知能学会論文誌, Vol.28, No.4, pp.370-378, 2013.

[2] 葦原史敏, 木村泰知, 荒木健治: 地方議会会議録における節単位による議員の要望抽出, 電子情報通信学会論文誌, Vol.J98-D, No.11, pp.1390-1401, 2015.

[3] 高丸圭一, 内田ゆず, 乙武北斗, 木村泰知 : 地方議会会議録コーパスにおけるオノマトペ -出現傾向と語義の分析-, 人工知能学会論文誌, Vol.30, No.1, pp.306-318, 2015.

[4] Yasutomo Kimura et al., Creating Japanese Political Corpus from Local Assembly Minutes of 47 Prefectures, Coling 2016, The 12th Workshop on Asian Language Resources, pp.78-85, 2016.

[5] 木村泰知, 小林暁雄, 坂地泰紀, 内田ゆず, 高丸圭一, 乙武北斗, 吉田光男, 川浦昭彦, 地方政治コーパス構築における従来の成果と現在の課題 -政治・経済分野の応用研究に向けたパネルデータの構築-, 言語処理学会第 23 回年次大会, C1-5, 2017.

[6] 齋藤誠, 大城卓, 菅原晃平, 永井隆広, 渋木英潔, 木村泰知, 森辰則: 地方議会会議録の収集とコーパスの構築, 言語処理学会第 17 回年次大会, P2-21, 2011.