Vol.2022-CH-129 No.2
2022/5/21

地方史統計資料における 知的探索インタラクションのためのデジタル情報化

中小路久美代[†1] 藤原慎太郎[†1] 寺沢憲吾[†1] 山本恭裕[†1]
川嶋稔夫[†1] 木村健一[†1] 松原伸人[†2]

概要：本研究が目指すのは、長期に渡って蓄積された多様な分野に渡る多数の統計表を対象として、データに包含される情報のインタラクティブな抽出、並置、統合、連携を可能とする新たなデジタルアーカイブ手法の構築である。人々が、地元のミュージアムを訪ねて地域の移り変わりを学んだり歴史的経緯の面白さに触れたりするのと同様に、市史統計表データとのインタラクションを通して知識の創出や仮説の生成を支える情報環境を創出したいと考えている。「函館市史：統計史料編」（函館市史編さん室(編): 函館市、 1987）は、昭和 62 年 5 月に発行された、明治初期から昭和後半にかけての函館市史に関わる 1285 ページに渡る統計資料表を編んだものである。気象、行政、財政、漁業、農業、工業、風俗、水道・電気・ガス、教育、兵事等に係る、計 317 個の表が収められている。2020 年より開始した StatsHakodate プロジェクトではこれまでに、この史料編の表が有する価値とそのデジタルデータ化による保存と展開を考察してきた。本稿では、印刷された統計表資料のデジタルデータ化技術の開発と、統計史料表とのデータエンゲージメントに向けた理論的考察および表データの液状化と結晶化を踏まえたデータアラインメントのデザインについて説明し、本プロジェクトの展望を論じる。

キーワード：市史統計表, 地方史, 表データエンゲージメント, インタラクションデザイン

Knowledge Interaction Design for a Large Amount of Tabular-based Local Statistical Historical Data

MKUMIYO NAKAKOJI[†1] SHINTARO FUJIWARA[†1] KENGO TERASAWA[†1]
YASUHIRO YAMAMOTO[†1] TOSHIO KAWASHIMA[†1] KEN-ICHI KIMURA[†1]
NOBUTO MATSUBARA[†2]

Abstract: Our goal is to build a new digital archiving method that enables interactive extraction, juxtaposition, integration, and linkage of information contained in the large volume of statistical table data that have been accumulated over a long period of time spanning a variety of fields. We aim at developing computational environments that support hypotheses-making and knowledge-building by interacting with the data of statistical tables of municipal history, in the same way that people visit a local museum to learn about the historical development of the region or to find out historical events that interest them. "Hakodate City History: Statistical Archives" published in May 1987 by City of Hakodate, is a compilation of 1,285 pages of statistical tables related to Hakodate City History from the early Meiji period to the late Showa period. The StatsHakodate project, launched in 2020, has been examining the value of the tables in this archive and how people could engage in the tables when stored as digital data. This paper describes the development of digital data conversion techniques for printed statistical table materials, theoretical considerations for data engagement with statistical historical tables, and the design of data alignment based on the liquefaction and crystallization of table data. The paper concludes with a discusses the prospects for this project.

Keywords: Statistical Tables, Local Histories, Tabular Data Engagement, Interaction Design

1. はじめに

本研究が目指すのは，長期に渡って蓄積された多様な分野に渡る多数の統計表とのインタラクションにより，データに包含される情報の抽出，並置，統合，連携による知識の創出や仮説の生成を支える情報環境の創出と，それを可能とする新たなデジタルアーカイブ手法の構築である．

「函館市史：統計史料編」（函館市史編さん室(編): 函館市史：統計史料編, 函館市, 1987）は，数百の統計表から成るデータから構成される膨大な情報空間である．明治初期から昭和後半にかけての函館市史に関わる 1285 ページに渡る統計資料表を編み，昭和 62 年 5 月に発行された．気象，行政，財政，漁業，農業，工業，風俗，水道・電気・ガス，教育，兵事等に係る，計 317 個の表が収められている．

次章では，本研究の目的および背景にある考え方を述べる．続いて，印刷された統計表資料のデジタルデータ化技術の開発と，統計史料表とのデータエンゲージメントに向けた理論的考察および表データの液状化と結晶化を踏まえたデータアラインメントのデザインについて説明する．最後に本プロジェクトの展望を論じる．

1 公立はこだて未来大学
Future University Hakodate
2 株式会社 SRA
Software Research Associates

2. 本研究の背景と目的：「函館市史：統計史料編」という情報空間が有する可能性

2020 年より開始した StatsHakodate プロジェクトでは，この史料編の表が有する価値とそのデジタルデータ化による保存と展開を考察してきた．「函館市史：統計史料編」に編まれた多数の統計史料を読み進めることにより，統計データとしての情報の価値に加えて，以下に述べる様々な面白さがあることが明らかとなった[1].

掲載されている統計表のそれぞれにおいて，統計データ，数値としての情報に加えて，表側にリストされている項目自体や，また該当する期間のデータの存在の有無そのものなどが，非常に興味深い．たとえば，明治 11 年と 12 年の「函館市中職業表」では，当時の函館市内の地域ごとに，細かな職業とその人数を記載しており，「五十集（いさば：漁師ではない魚を扱う）」や「立花師（たてはなし：生け花をする）」といった，現在では使用しなくなった用語も多く出現する．

また，表側内のインデント（字下がり）や罫線割り，ラベルを用いた構造化が多数用いられており，表計算ソフトウェアを利用する以前の表の組版としての面白さもある．史料編の中に現れる表の大きさと丁付け（ページ割り）の関係も興味深い．さらに，含まれている数百枚の表を見比べると，表形式内に現れる配置や間隔（スペース）の取り方に多様性があることに気づく．罫線と行間の取り方が異なるものがみられ，恐らく組版を組んだ人が章によって異なったと考えられる．

さらに，函館市史：統計史料編で取り扱われている歴史的な統計資料データに関わる多様な時間軸が，重層的に現れる点も興味深い．データが対応する時間に加えて，それが記録された元資料の編集あるいは発行時期，さらに本統計史料編として編まれた時期，という三つの時間が重なる．「凡例」によると，函館市史：統計史料編が主として利用したものとして，開拓使事業報告，北海道庁統計書，北海道統計書，函館市長統計概表，など，計 18 個の文献がリストされている．「部門別解説」の中には，それぞれの項目の統計表の来歴や使用した資料，またその考え方などが記載されており，発行時期による調査方法の違いなどといった時間的背景を知ることができれば，表データをより深く解釈できるとも考えられる．

StatsHakodate プロジェクトで取り組むのは，数百の表のみから構成される統計資料を対象として，表データそのものを対象とするデジタルアーカイビングの試みである．人々が，地元のミュージアムを訪ねて地域の移り変わりを学んだり歴史的経緯の面白さに触れたりするのと同様に，市史統計表データとのインタラクションを通して知識の創出や仮説の生成を支える情報環境を創出したいと考えている．ミュージアムは，人々を啓蒙するという従来の機能から，人々が楽しめる場，触発する体験を得るための場を提供する空間へと，その役割を変換しつつある[2][3]．函館市史：統計史料編をデジタルアーカイビングするにあたって，市史統計表データから構成されるインタラクティブな情報空間を，触発する体験を可能とするような文化的な場としてのミュージアムのように構成し実現したい．

表は，人類が生み出したビジュアルな情報表現のひとつの形態であると考えられる．セルと呼ぶ升目を水平方向（行）および垂直縦方向（列）に２次元に整列配置し，行および列のそれぞれに属性を与える．行と列の交わる部分に，両者の属性に対応する数値や文字といったデータが配置される．表を見る際には，行および列の位置から水平方向，垂直方向に辿ることで，値を求めたり，ある行内やある列内を順に見ていったりすることで，その変化の傾向を知ることができる．

我々は，データに内包される情報を読み取るに留まらず，データに対する多様な視覚的インタラクティビティを介してデータの背後にある世界に思いを馳せ，知識創出につながるようなデータへの関わり方を，データエンゲージメントと呼ぶ[4]．本研究で狙うのは，表として印刷された歴史的な紙媒体を，発展的，可塑的な形でデータ化し，利用者がその体験を通して新たな課題を発見したり，創造性を発揮したりできるようなデータエンゲージメントのための情報環境の創出である（図 1）．「函館市史：統計史料編」という具体的な資料を対象として研究を進めることで，広く，国内外の自治体の地域史など，表形式で残されている知識媒体への適用を目指す．

図 1 創造や触発を促すデータエンゲージメントのための
データインタラクティビティ

3. データ化技術の開発

StatsHakodate プロジェクトではこれまでに，物理的な紙のページに印刷された表を，デジタルデータとして読み取り，csv ファイルといった表形式データとして扱えるようになるデータ化技術の開発に着手した．表は，表頭，表側，表体といったパーツから構成される．デジタルデータとしては，これらの表構造を認識した上で，各部に記載された文字が認識できることが求められる．

市販の文字認識ソフトウェアを用いて，本プロジェクトで事例として取り組む「函館市史 : 統計史料編」の冒頭の日本語自然文記述のページ画像に対して文字認識させると，99 パーセントを超える認識率で認識できることがわかった．しかしながら，表を記載したページのスキャン画像の読み取りでは，表頭や表側といった表の構造認識において誤認識が頻発すること，行（横方向）と列（縦方向）の対応においても誤認識が生じ易いこと，という２点が明らかとなった．

「函館市史 : 統計史料編」では，その表体において，行ごと，列ごとの罫線をもたない表が多い．一般に，人が紙面上で読み取ることを前提として印刷された表は，ページ上の視覚的な煩雑さを避け可読性の向上を図るために，罫線を引かないことが提唱されていたとされる．さらに，「函館市史 : 統計史料編」においては，ひとつの表が 3 ページ以上の複数ページに渡って掲載されているものに加え，1 ページ内に複数の表を掲載するものや，ページ内で左方向に 90 度回転させて横向きに表を掲載しているものもある．

本プロジェクトにおいては，市販の文字認識ソフトウェア（読み取り革命 Ver.16）を用いて文字認識をおこなうこととし，その認識精度を向上させるために，紙面をスキャンした画像データに，画像処理を施すこととした（図 2）[5].

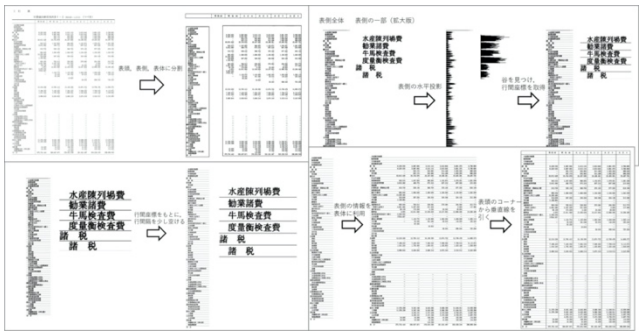

図 2 文字認識精度向上のための画像処理（[3]より抜粋）

具体的には，「函館市史 : 統計史料編」における表では表頭上部の水平線が他の罫線よりもやや太い線で示されていることを利用して，表の向き，および表構造の検出をおこなった．表側および表頭内の文字の印字量の多寡を用いて行間箇所を抽出した上で行間を広げたり，罫線との重なりを配慮して横方向に文字をずらしたりして，表のセル同士や行間の空白を広げた．試行した画像処理は，表の向きの認識と修正，表の傾き補正，セル内の文字表示位置の補正，罫線の修正（削除と追加），新たな罫線の追加，などである．

「函館市史 : 統計史料編」からランダムにサンプリングしたデータを用いて，追加した罫線と文字列の重なりを調査した結果，50141 個のセルに対して罫線と文字が重なってしまっているところが 207 箇所あることがわかった．また，これらの重複箇所は，均等に分散するというよりはむしろ，いくつかの表にまとまって現れることがわかった．

罫線を追加するなど画像処理を施した，表の特徴がばらばらの 5 枚分の二値化画像に対する文字認識ソフトウェアの文字認識率は，98.25 パーセントであった．誤認識の事例を具体的にみてみると，文字間の余白をあけることで対処できそうな場合や，三点リーダー記号など表に頻出する文字への対処が必要なことが示唆された．

「函館市史 : 統計史料編」に掲載されている表の中には，表側や表頭の中で構造化されているものが散見される．罫線で明示的な構造化されているものもあれば，字下げを用いて構造を示しているものもある．今後は，罫線と文字との距離の把握など，字下げの有無やその程度の認識が必要となると考えられる．さらに，「函館市史 : 統計史料編」を具体的な対象として開発したデジタルデータ化技術を，他地域の統計史料など，より一般的な表形式画像の読み取りに適用していくことが考えられる．ページスキャン画像におけるページの汚れやページのよれなどといった修復技法を取り入れていくことも考えられる．

4. 知的探索インタラクションのための情報環境

膨大な表形式のデータを，人が理解しやすい形で表現するにあたっては，データをビジュアルに図示することが考えられる．人は，極めて少ない認知負荷で，視覚的表現間の比較や順序づけを短時間に行える．Diagram（図表）による表現は，文章による表現と比して，(情報の種類によっては) 認知的処理において格段に効果的であるとされる[6] ．

本プロジェクトでは，我々がこれまでに培ってきた，データサイエンスにおける大規模なデータ可視化におけるデータインタラクティビティの研究成果を取り入れることとした．すなわち，データそのものの理解ではなく，データの背後にある世界を理解するような，創造や触発を促すデータエンゲージメントのためのデータインタラクティビティである[4]．表現 (representation) の有効性は，表現それ自体にあるのではなく，表現を利用しそこから推論する操作者の側にあるとされる[6]．データサイエンスにおける大規模なデータのための視覚的表現は，それを受け取る利用者が，その表現からいかに効果的に仮説や知識を生成し，検証することができるか，という点からその有効性が論じられるべきものである．

データサイエンスにおいて扱われる大規模データを視覚的に表現するにあたっては，伝統的なグラフやダイアグラム表現とは異なる，三つの課題がある [4]．第一に，対象とするデータについて，その示唆する意味や価値といったものが，あらかじめ自明ではないという課題がある．第二に，コンピュータディスプレイ上で表示される視覚的表現は，人がインタラクティブに指示する軸や視点の変更と連動する形で，極めてダイナミックに変化し得るという点がある．

データの視覚的要素への符号化に加えて，人が操作する対象を表す視覚的要素を，可視化表現に組み込み，連動させる必要がある．第三の課題は，データに対する人の関心や興味は，個々人が有する関連する知識や体験に大きく依存すること，さらに，視覚的表現との対話を通して関心や興味が時間的に遷移する点にある．表示された視覚的表現を見ることで，人は新たに，注目すべき点や観点に気づく．その結果を踏まえて，さらに別の視点からデータの視覚的表現を求めるといったことが繰り返される．このサイクルは，非常に短い時間で繰り返されることもあれば，人の問題解決の思考の過程に応じて時間をかけて繰り返される場合もある．同じ視覚的表現であっても，誰がその表現を見て，どの経路でその表示に辿り着いたのかによって，それぞれの人の解釈は異なってくると考えられる．

本プロジェクトでは，「函館市史：統計史料編」の膨大な統計資料表を，インタラクティブな可視化表現を介して体験することで，統計資料が収集，整備された背景にある世界の理解につなげたいと考えている．本プロジェクトが拠り所とする仮説は，統計資料表をデジタルデータ化した上で，多数の表データから成る膨大な情報空間とのデータインタラクティビティを実現することで，地域発展の経緯や歴史に興味や関心を抱くといった体験の広がりを提供することができるというものである．

5. 統計史料表データエンゲージメントに向けたインタラクティビティの要素

統計史料表データの背後にある世界や当時の状況に思いを馳せ，触発と創造を促すような革新的なデジタルアーカイブ環境の実現に向けて，必要となる表データとのデータインタラクティビティの要素を考察した．

これを可能とするため本研究では，「知識の液状化と結晶化」の手法[7]を，統計資料表に対して適用する．表というデータの表現形式を，いったん，行ごと，列ごとといった多様な形式を有する素材として扱い，それらを多様な文脈で自在に組み合わせることで，新たな発見や体験を可能とすることを目指す．

統計資料表は，それが整えられた時点で観察された事項や得られた情報を，その時点でわかりやすい形態に表現したものである．このように形式的に表現された知識は「知識の一形態」にすぎず，知識とは「そうした『個体』のようなものではなく『液体』のようなもの」であり，「文脈によって形を変えることや，部分的に抽出して融合することで新たな文脈に適用可能な性質を持つものに変化させることができる」ものである[8]．知識の液状化では，ある文脈に沿って静的に構造化され表現された知識表現を，液体のように自由な形で見てとれるようにする．ある文脈でその構造を見てとるような表現として表すことが，知識の結晶化である．

知識として表された情報を「液状化」するための技術的なアプローチとして考えられるのは，静的に表現された知識に多様な文脈を与え，知識表現を構成する部分間に多数のつながりを与えることによって，構造を軟化させることができる．部分間の関連性（リンク）を断ち切るのではなく，逆に多数与えることによって，構造に可塑性を与えるアプローチである．

「函館市史：統計史料編」に掲載されている多数の表を対象として，これらの液状化と結晶化をおこなうにあたって，これまでに以下 10 個の方式を検討している．本プロジェクトにおける結晶化のプロセスは，データを揃えて見せること（データアラインメント[4]）とする．

- 表体を，表側との関連を保ったまま列ごとに分離する
- 表体を，表頭との関連を保ったまま行ごとに分離する
- 表側に現れる時間情報（主として「年号」表記）で列同士を連携する
- 表頭に現れる時間情報（主として「年号」表記）で行同士を連携する
- 表側に現れる時間情報と表側に現れる時間情報で行と列を連携する
- 表頭あるいは表側に現れる地名で連携する
- 表頭あるいは表側に現れる地名の座標位置で連携する
- 表頭あるいは表側に現れる時間情報を用いて外部情報と連携する
- 表頭あるいは表側に現れる地名を用いて外部情報と連携する
- 表セル内に現れる名称で外部情報と連携する

時間情報を用いた連携においては，年号表記と西暦表記のマッピングに加え，離散的な時間表記と，時間幅をもった連続的な時間表記との連携のさせ方を考える必要がある．また，地名を用いた連携においては，これまでに，京都市の歴史年表とウィキペディアを用いて地名の経緯度を取得する方法を適用することを考えている．さらに，年代により地名が変更されていたり，同じ地名でも地理的な行政域が異なっていたりする場合も考えられ，これらを連携させる方式の構築が必要となる．

外部情報との連携としては，ひとつには，他地域の統計情報との連携が考えられる．たとえば福井県は，統計資料をウェブページで公開している．これらの公開されているデータと連携して，たとえば同じ時期の函館市の人口の推移と並置してみるといった単純な可視化から始めて，より重層的な表データの「結晶化」を目指す．

名称を用いた外部情報との連携としては，たとえば項目中に現れる語句の意味を，ウェブ上のデータから検索して提示するといったことが考えられる．たとえば明治初期の職業表では，今では馴染みのない職業名も多い．それが具体的に何を指しているのかを，表データの体験と隣接した形で探索することで，表データへの興味が増していくと考え

られる．

これらの液状化と結晶化の方式を用いて，データインタラクティビティにおけるデータアラインメントを可能とするためには，各表が有する表側，表頭および表のタイトルから類推される，時間的，地理的連携可能性をあらかじめ求めておく必要がある．ユーザの求めに応じて動的に連携させていくための内部的なデータ表現形式のデザインは今後の課題である．

6. 考察

IIIF (International Image Interoperability Framework) Manifest で表現されたデジタルアーカイブデータに向けて，それを編集したりブラウジングしたりするためのツールが公開されているが，表という表現形式を対象とした IIIF 表現の例は，筆者らの気づく限りにおいては国内外を通して未だ公開されていないと思われる．「SAT 大蔵経 DB」プロジェクト[9]や「デジタル法寶義林」プロジェクト[10]などの既存のデジタルアーカイブプロジェクトのアプローチの多くは，アーカイブされたデータへのアノテーションを通して，対象となるデータを繙くことが目的である．これに対して本プロジェクトでは，史料を素材とした様々な可視化を通して，専門家のみならず市民や学生などより多くの人々が，その背景にある地域の歴史や発展に思いを馳せ，編み上げていくことができるような環境を目指すものである．

今後は，本研究で構築する情報環境を国内外の自治体の地域史など，表形式で残されている知識媒体へ広く適用したいと考えている．たとえば福井県は統計表データを公開しているが，ある時期までの紙媒体のデータと，デジタルに収集されたと思われる統計データとを分けて公開しており，前者ではページをスキャンした PDF 形式の画像データの公開に留まっている．本プロジェクトで開発したデジタルデータ化技術を，既存の公開されている表画像データに適用することも検討中である．

本プロジェクトの成果の一部として，デジタル化したデータの一部を，函館市と連携して公開することを検討している．今後は，スキャンした画像データの公開を端緒に，市民をはじめデジタルデータの活用を可能とするよう，展開を目指す．

謝辞 本研究は，公立はこだて未来大学特別研究費（重点領域）の支援を受けたものです．「函館市史：統計史料編」の研究活用においては函館市の協力を感謝します．

参考文献

[1] 中小路久美代，山本恭裕，松原伸人，川嶋稔夫，木村健一，函館市史:統計史料編のデジタルデータ化における多角的検討，情報知識学会誌, Vol.30, No.2, pp.176-181，情報知識学会第 28 回年次大会, 2020.
[2] 中小路久美代，新藤浩伸，山本恭裕，岡田猛（編），触発するミュージアムー文化的公共空間の新たな可能性を求めて，あいり出版、May, 2016.
[3] 中小路久美代，岡田猛，川嶋稔夫，山本恭裕，新藤浩伸，木村健一，影浦峡，文化的な公共空間における触発する体験，サービソロージー, Vol.3, No.2，サービス学会, pp.10-17, July, 2016.
[4] 中小路久美代，山本恭裕，松原伸人，北雄介，データ可視化におけるデータインタラクティビティ，電子情報通信学会誌，「データサイエンスにおけるデータ抽象化によるデータ理解へのアプローチ」小特集号, pp.197-205, 2021.
[5] 藤原慎太郎，印刷された統計資料表のテキスト化技術の開発，卒業論文，公立はこだて未来大学システム情報科学部, 2022.
[6] Larkin, J.H., Simon, H.A., Why a Diagram is (Sometimes) Worth Ten Thousand Words, Cognitive Science, Vol.11, No.1, pp.65-100, Wiley, March 1987.
[7] Hori, K., Nakakoji, K., Yamamoto, Y. and Ostwald, J.: Organic Perspectives of Knowledge Management: Knowledge Evolution through a Cycle of Knowledge Liquidization and Crystallization, Journal of Universal Computer Science, Vol.10, No.3, pp.252–261, 2004.
[8] 網谷重紀，堀浩一，知識創造過程を支援するための方法とシステムの研究，情報処理学会論文誌, Vol.46, No.1, pp.89-102, 2005.
[9] SAT 大蔵経 DB, https://21dzk.l.u-tokyo.ac.jp/SAT/tutorials.html
[10] 渡邉要一郎，永崎研宣，大向一輝，井野雅文，村瀬友洋，朴賢珍，下田正弘，デジタル法寶義林における研究データの共同構築，情報処理学会，研究報告人文科学とコンピュータ, 2022-CH-128, pp.1-4, 2022.