第28回年次大会予稿

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

Toward Digitalizing Historical Statistical Tables of Hakodate

中小路久美代^{1*}, 山本恭裕¹, 松原伸人², 川嶋稔夫¹, 木村健一¹

Kumiyo NAKAKOJI^{1*}, Yasuhiro YAMAMOTO¹, Nobuto MATSUBARA²,
Toshio KAWASHIMA¹, Ken-ichi KIMURA¹

1 公立はこだて未来大学

Future University Hakodate

〒041-8655 函館市亀田中野町116-2

2 株式会社SRA

Software Research Associates Inc.

〒171-8513 東京都豊島区南池袋2-32-8

*連絡先著者 Corresponding Author

函館市史：統計史料編は、昭和62年に発行された、明治初期から昭和後半にかけての函館市史に関わる1200余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、数百余の表が収められている。本稿では、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する。これまでの検討から、表が有している歴史的 content やそこから類推される事象に加えて、表という形式のデジタルデータ化の手法、表を書籍として出版するという印刷メディアとしての制約やコンベンションの同定、人手による誤植や誤記入の取り扱い、グラフィックデザインとしての作法の表現など、考慮すべき点が明らかとなった。本稿では、これらの観点から、函館市史統計資料編をデジタル化の課題を考察する。

The Volume of Historical Statistical Tables on Hakodate City was published in 1987, which is the compilation of more than 1,200 historical data tables, where the data covers the period between 1870's and 1970's. The topics of data are composed of climate, politics, finance, fishery, agriculture, industry, culture, education, social infrastructure, and military. This paper examines how we could engage in such a large number of historical data tables from various areas, and discusses the technical issues and research challenges that the digitization of the historical resources would raise.

キーワード: 統計史料, 表形式データ, 統計史料のデジタルアーカイブ

Keywords: Historical data, Table-based historical data, Digital archives of historical statistics

1 はじめに

本論では、函館市史：統計史料編に含まれる、表形式で表現された大規模データのデジタルアーカイブについて考察する。函館市史：統計史料編は、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（図1）。たとえば図1右に示すような、計300個あまりの表が収められている。数ページにわたって分割された表も多い。

Figure 1 shows the cover of the book 'Hakodate City History: Statistical Materials' (函館市史：統計史料編) on the left. On the right, a sample table is displayed, showing population statistics for various districts in Hakodate. The table is organized into columns for district names, population counts, and other statistical data.

図1：函館市史：統計史料編（左：本編写真，右：表の例（[1, p. 474]）

我々は、本書を構成する大量の統計表を対象として、それらを多様な観点から吟味しデータ空間を探索するようなインタラクションを可能とするデジタル化を目指している。これまでに、本書の全ページをスキャンし、PDFおよびJPEGデータとして保存した上で、ページ内の表位置の認識や、表構造の抽出に着手したところである。次章では、函館市史：統計史料編についての概要と、表構造を取り扱う上での用語を説明する。続いて、表データを我々自身が閲覧しながら気づいたことや興味深い点を列挙し、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する。

2. 対象とする史料

2.1 函館市史：統計史料編について

函館市史：統計史料編の本文は、冒頭の「部門別解説」に続いて、それに続く「行政」（5章）、「産業・経済」（10章）、「社会・教育」（6章）という三部構成の本体、および巻末の付表（「戦没者名簿」）から構成されている。部門別解説には、それぞれの部門の統計表の来歴や、データ収集における考え方、また項目名や判例の説明の記載がある。各章は、それぞれ当該項目についての様々な統計表で構成され、計21章から成る（表1）。本体および付表の各ページには文章はなく、全て表が収められている。

表1：函館市史：統計史料編の本体の構成と取り扱われている分野

I 行政	II 産業・経済	III 社会・教育
1. 気象	1. 農業	1. 風俗
2. 土地	2. 漁業	2. 消防
3. 人口	3. 露宿漁業	3. 衛生
4. 財政	4. 工業	4. 水道・電気・ガス
5. 選挙	5. 事業所	5. 教育
	6. 商業・金融	6. 兵事
	7. 運輸・通信	
	8. 内国貿易	
	9. 物価・資金	
	10. 司法・警察	

2.2 表を構成する各部の名称

本研究では、大量の表形式で表現された情報を取り扱うにあたり、図2に示す各部の名称を用いる。

図2：表を構成する各部の名称

いわゆる罫線で囲まれた表本体として我々が認識するのは、(b)から(d)の矩形で囲まれた領域である。この内容を補足する情報が、(f)(g)(h)となる。いわゆる注意書きは(g)の脚注に記載される。(h)の頭注には、単位などが記載されることが多い。(i)および(j)は、表を出版物として製本する上で現れる要素である。(a)の表題目は、出版物の目次や表番号として整理される際に用いられる。

3. 函館市史：統計史料編の面白さ

本節では、函館市史：統計史料編を閲覧しながら我々が気づいた興味深い点を解説する。

3.1 表側内容の面白さ

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。

「職業表 1-1~2」([1; p. 88-91])を例にとる。この表は、明治6年から明治25年までの間の、職業とこれに従事する人数を毎年列挙したものである。職業は「農・工・商・雑」に分類され、農の中には「薪炭」や「木挽」といった職業がリストされている。雑には、「官吏」「産婆」「芸妓」「船乗り」などがみられる。

明治11年と12年に関しては「函館市中職業表」がある。これは、当時の函館市内の地域ごとに、より細かな職業とその人数を記載するものである。「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。

職業表に関してはこれとは別に、「居留

外国人職業表が」2種類あり、明治5年から明治34年、大正14年から昭和5年のデータである。また、明治25年以降の職業表が掲載されていないのも興味深い。

このように、おおよそ100年に渡る統計史料を読み解くにあたっては、数値の変遷はもちろんのこと、項目の変化や項目中に現れる語彙、データの存在の有無の変化や変遷、といった事柄にも着目できることが望ましいと考えられる。

3.2 表側内の構造

函館市史：統計史料編をデジタル化するにあたって大きな課題となると考えられる事柄のひとつに、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ある。

図1右に示すのは、「漁業経営体・漁業者」に関する表の表側の一部である。項目間の構造関係が、表側の中に一部表形式として表現されている。縦の罫線でグループ化がされている部分と、そうでない部分がある。また、「A」「B」「C」「D」といったラベルが付与されており、その合計値の項目もある。縦書きと横書きが混在しており、判別が明らかでないところもある。当該分野の知識をもたない人にとっては、意味を解釈しながら読み解くことも難しく感じられる。

3.3 表の丁付け

表を印刷物として発行する場合、表の大きさによってはページをまたがらせて、表を分割する必要がある。これは丁付け（ページ割り）と呼ばれる。函館市史：統計史料編の中に現れる表の大きさと丁付けの関係を、図3に示す。縦横ともに1ページに収まらない場合は、横方向が優先され連続したページに表示されている。また、表側

の項目数が極めて少なく縦に短い表の場合は、ページ内で折り返して表示されることがある(e)。同様に、複数の異なる表が1ページに収められているものも1箇所ある。

図3: 表の大きさと丁付けのパターン

また、書籍として印刷されている函館市史：統計史料編には見開きがある。横書きの本書では、偶数ページが左側、奇数ページが右側にある。ページが見開きの左右どちら側であるかによって、表側頭および表側の掲載の仕方が異なる。(c)や(d)の場合、連続したページが見開きの左側（偶数ページ）となる場合は、表側頭および表側が左端に再掲される。見開きの右側（奇数ページ）となる場合は、これらは省略される。デジタル化においてページ内の表構造を自動認識するにあたっては、分割された表の再構成が必要となると共に、見開きページの左右の情報も用いながら、表側の有無を推定することになると考えられる。

3.4 人手による誤植や誤記入の取り扱い

いくつかの表においては、明らかな誤植、さらには、データを捏造あるいは改ざんしたのではないかと考えられる箇所が見つかった。

後者の事例として図4に示すのは、「明治期町別戸口表」の一部である。個別表は、年ごとに、各地域の個数と人口を列挙したもので、それぞれに一戸平均が算出されている。

ここで、明治33年の列をみると、すべての一戸平均が、5.00となっている。恐らく、戸数に5.00を乗算してその値を人口値として列挙したものと推測される。この年だけは人口を調査しなかったのか、あるいは調査値が紛失されたのかといったことが思い浮かぶ。他の一戸平均数値をみると、その値はおおむね3から4の間であり、なぜ「5.00」という数値を用いているのかといったことにも問いが生じ、興味深い。

明治 30	一戸	明治 33	一戸	明治 36	一戸	明治 38	一戸			
戸数	人口	戸数	人口	戸数	人口	戸数	人口			
90	297	3.30	70	350	5.00	123	480	3.90	142	534
162	685	4.23	169	845	5.00	267	1,269	4.75	195	1,280
157	596	3.80	132	660	5.00	256	808	3.15	236	891
147	400	2.72	162	810	5.00	266	1,032	3.88	291	1,011
358	1,127	3.15	354	1,770	5.00	625	2,326	3.72	654	2,400
433	1,456	3.36	430	2,150	5.00	990	3,120	3.15	776	2,330
						32	112	3.50	24	139
						36	121	3.36	110	484
						29	124	4.28	75	282
						13	48	3.69	28	120
						62	261	4.21	98	417
308	1,012	3.29	314	1,570	5.00	555	1,852	3.34	583	1,927
289	929	3.21	251	1,255	5.00	443	1,423	3.21	425	1,449
326	1,330	4.08	533	2,665	5.00	832	2,847	3.42	876	3,077
153	494	3.23	145	725	5.00	223	645	2.89	226	777
238	914	3.84	283	1,415	5.00	513	1,927	3.76	322	1,840
192	700	3.65	205	1,025	5.00	327	1,289	3.94	305	1,213
75	117	1.56	75	375	5.00	117	439	3.75	102	407
71	279	3.93	66	330	5.00	109	437	4.01	96	392

図4: 明治期町別戸口表から一部抜粋
([1; p. 88])

3.5 グラフィックデザインとしての面白み

函館市史：統計史料編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ていたと考えられる。図xxに示すように、文字間の行間と文字と罫線の間の行間が等しく罫線が引かれているものと、文字間の行間は等しく配置され、罫線はあとから足されたように見受けられるものがある。

また表の中には、表側中の項目10個ごとに空行があるものがある。具体的な項目の内容を確認してもこれらに連関性はなく、空行が10個ごとに繰り返すことを踏まえると、この空行は項目のグルーピングのためではなく、恐らく見易さのために区切り

として挿入されていると考えられる。また、誤って11個目に空行がはいっている箇所もある。意図的なものか意図せず挿入された印刷ミスであるかの見極めは、人がじっくり見ても難しい。

函館市史：統計史料編の発行時期には、パソコンや表計算ソフトはまだ広く普及はしておらず、資料の収集および印刷製本には、多くの手作業が関わっていたと考えられる。活版印刷や写真植字といった当時の技術や手法そのものが示唆する文化的な興味深さも見てとることができる。

3.6 重層的な時間の取り扱い

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をアーカイブする際に興味深い点のひとつとして、多様な時間軸が重層的に現れる点がある。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。これらに加えて、「部門別解説」の中に、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いる。このような、元とした資料の発刊時期は本書には記載がないものも多い。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。

4. デジタル化に向けての考察

我々は、統計史料編を構成する大量の統計表を対象として、それらを多様な観点か

ら吟味しデータ空間を探索するようなインタラクションを可能とするデジタル化を目指している。このような多数の統計表を対象とした構造分析や意味解析のアプローチは、我々の知る限りにおいては、これまでにあまり見られないように思われる。

表の構造解析や、表の意味構造の理解についての既存の情報科学的アプローチの多くは、PDFで表現された学術論文などのドキュメント中から、機械学習によって表構造を認識するものや[2][4]、請求書から表構造を抽出するもの[5]である。表構造の認識結果を用いて、その表の内容をグラフとして自動生成するアプローチもある[7]。

また、ウェブページを対象として、HTMLの<table>タグを用いて記載された行列の順序付きセル集合を認識し、その自動分類を行う研究や[3]、表の意味構造を理解する深層学習を用いたアプローチも見られる[6]。

これら既存研究はいずれも、数個から十数個といった個数の表の構造抽出と意味理解を目指すものであり、あくまでも表はドキュメントを理解するための補助的情報として位置づけられているように捉えられる。また、その多くがデジタルデータとして生成された表を対象としている。これに対して我々が目指すのは、アナログに生成された、表が主役のデジタル情報化である。ほとんどのページが表のみから構成され、ページ中から表構造を抽出する課題自体は比較的取り組み易い。同時に、アナログに生成されていることから生じる、人為的な罫線の上書きといった課題が生じる。

我々が対象としているのは、個別の表の取り扱いではない。函館市の歴史という一貫したテーマに関する多岐な分野にわたる多数の表である。個々の表に現れる項目間の関係性や、異なる表にまたがる関連項目の連携を、並べ替えやフィルタリングといったインタラクションを通してみたいと考えている。中でも函館市の史料における連携手法として最も重要なことは、(1)時間情報による連携、(2)地理情報による連携、および(3)人名や会社名、組織名といった固有名詞による連携であろうと考えている。

さらには、前節で掲げたような、史料としての表側内容の面白さ、表側内の構造の取り扱い、表の丁付けの取り扱い、誤植や誤記入の取り扱い、グラフィックデザインとしての面白み、函館市史：統計史料編という書籍の存在を取り巻く重層的な時間構造、といった事柄を踏まえることによって、歴史家はもちろんのこと、一般市民や地域に興味をもつ人々が楽しみ触発されるようなデジタルリソースとして活用したいと考えている。

今後は、函館市史：統計史料編に特化した表構造の認識技術の構築を進めることと並行して、表構造の抽出をマニュアルで試行し、既存の研究手法を用いた数値のグラフ化を行うなど、連携化の試作を行っていく予定である。

参考文献

[1] 函館市史編さん室(編)：函館市史：統計史料編，函館市，1987.

[2] F.F. Babatunde, B.A. Ojokoh, S. A. Oluwadare: Automatic Table Recognition and Extraction from Heterogeneous Documents, Journal of Computer and Communications, Vol. 13, no. 12, pp. 100-110, 2015.

[3] E. Crestan, P. Pantel: A fine-grained Taxonomy of Tables on the Web, Proceedings of CIKM 2010, pp. 1405- 1408, 2010.

[4] S.F. Rashid, A. Akmal, M. Adnan, A.A. Aslam, A. Dengel: Table Recognition in Heterogeneous Documents Using Machine Learning, Proceedings of ICDAR 2017, pp. 777-782, 2017.

[5] P. Riba, A. Dutta, L. Goldmann, A. Fornés, O. Ramos, J. Lladós: Table Detection in Invoice Documents by Graph Neural Networks, Proceedings of ICDAR 2019, pp. 122-127, 2019.

[6] 西田京介, 貞光九月, 東中竜一郎, 松尾義博: TabNet: 表の意味構造を理解するハイブリッド型深層ニューラルネットワーク, DEIM (Data Engineering and Information Management) フォーラム 2017, B6-4, pp. 1-8, 2017.

[7] 山田凌也, 太田学, 高須淳宏: グラフの自動生成のための表の構造解析の一手法, (Data Engineering and Information Management) フォーラム 2019, E8-5, pp. 1-8, 2019.