e-Statでの統計データ検索におけるいくつかの課題抽出とその解決方法の提案

Method for improving the Recall in e-Stat Data Search

芦澤颯太† 松田純一 大曽根匡†
Souta ASHIZAWA† Junichi MATSUDA Tadashi OSONE†
†専修大学大学院　経営学研究科
† Graduate School of Business Administration, Senshu University.

要旨

e-Stat とは，政府統計の総合窓口のことであり，各府省等が実施する統計調査の各種情報を一つにまとめ，統計データの検索をはじめとした，さまざまな機能を備えたポータルサイトである．そして，政府が発信する統計データの源泉として多くの研究者や個人・企業などに活用されている．しかし，実際に使用してみると，検索漏れなどの課題があることがわかった．本研究ではその課題を具体的に抽出し，その解決方法について提案する．

1. はじめに

近年，IT の発展により，データサイエンスの活用が叫ばれてきている．そこで注目を集めているのがオープンデータである．オープンデータとは，デジタル庁が公開しているオープンデータの基本指針[1]によれば，「国，地方公共団体及び事業者が保有する官民データのうち，国民誰もがインターネット等を通じて容易に利用（加工，編集，再配布等）できるよう，次のいずれの項目にも該当する形で公開されたデータをオープンデータと定義する」としている．いずれの項目として，①営利目的，非営利目的を問わず二次利用可能なルールが適用されたもの，②機械判読に適したもの，③無償で利用できるものの 3 点を挙げている．

日本におけるオープンデータは，都道府県や市区町村などが公開しているほか，各省庁が公表している統計データをまとめて閲覧することができる e-Stat[2]という政府統計のまとめサイトが国によって公開され，広く利用されている．e-Stat は，2008 年より運用が開始された．それ以前は，各省庁が独自に統計データを管理し，情報提供を行ってきたので，利用者は省庁毎にアクセスが必要であった．それが，e-Stat のサービス提供により，各省庁の検索機能の重複排除や利用者への効率的な情報提供が可能になり，利用者の利便性が大幅に改善された．しかし，各省庁が別々に管理していたデータを 1 つにまとめたことによる影響や，日本経済新聞の記事[3]が指摘している省庁間の縦割り体質などによって，現在でもいくつかの課題がある．

2. e-Stat の管理構造

e-Stat は，現在約 30 の府省庁から集められた約 690 の調査を閲覧することが可能になっている．主な機能として，検索機能と活用機能がある．検索機能では，分野，組織，キーワードで検索が可能になっており，分野では，「人口・世帯」や「労働・賃金」など 17 の分野から，組織では「総務省」や「文部科学省」など約 30 の府省庁の統計データを利用できる．キーワード検索を用いて，利用者の目的に合ったキーワードで全文検索できる．活用機能は，グラフ，時系列表，地図，地域の 4 つの方法で統計データを表示できる機能である．

統計データは，階層的に管理されている．その統計データの管理構造を図 1 に示す．約 690 の調査のそれぞれの「統計調査&統計概要」の下位に調査年や都道府県別などの「提供分類」がある．そして，その下位に「統計表」がある．統計表の下位に「事項名」があり，事項名の下位にデータの名称を表す「項目名」がある．統計表の具体的な管理構造の実例を図 2 に示す．これは，令和 2 年の国勢調査における統計表「男女別人口」を簡略化したものである．図 2 の事項名「男女」では，下位概念として項目名「総数」「男」「女」がある．この事項名と項目名の階層構造を，図 3 のように表記することにする．性別に関しては，図 4 に示すように，事項名には性別に関係するようなワードがないのに，項目名には「プロ

グラマー（男）」ように性別に関係する名称になっていることも数多く存在する．また，図 5 のように，項目名が同じ「男」「女」の場合でも，統計表により，その上位概念の事項名が「性別」であったり，「男女別」であったりして，統一されていないことも少なくない．

統計表の具体例を，図 6 に示す．表の上部に「表頭項目名」が表示され，表の左側に「表側項目名」が表示される．このように項目名は統計表内に記載されているが，その上位概念である事項名は統計表に表示されない．したがって，事項名はユーザからは意識されないことが多い．特別な操作をしないと，項目名「総数」「男」「女」の上位概念の事項名が「男女」であることがわからない．

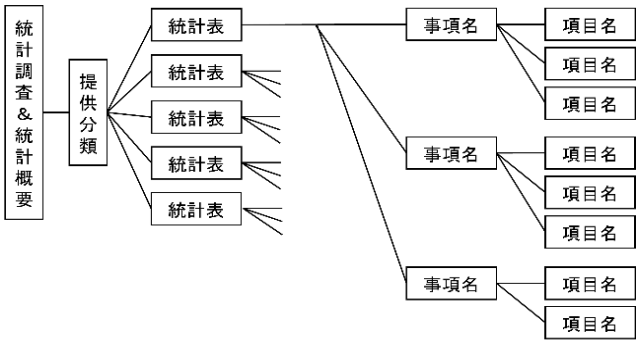

図 1　統計データの管理構造

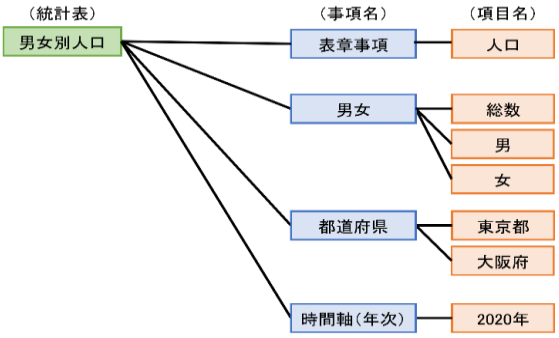

図 2　統計表の具体的な管理構造の例

男女	総数
	男
	女

図 3　事項名と項目名の表記

職種_128職種区分	プログラマー（男）
	プログラマー（女）

図 4　事項名「職種_128 職種区分」の項目名

性別	男
	女

男女別	男
	女

図 5　事項名の不統一の例

図 6　統計表の例

3. e-Stat 使用上の課題抽出

e-Stat 使用上の課題には，使い勝手の観点と，検索の正確性の観点からの課題があると考えた.

3.1. 使い勝手の観点からの課題

使い勝手の観点からの課題として，以下の 2 点が挙げられる.

(1) 統計表の絞り込みの困難さ

e-Stat には膨大な量の統計表が蓄積されているため，検索をすると多くの統計表がヒットしてしまうことが多い．絞り込みを行っても，数千件ものヒット数になることもあり，目的の統計表を数件に絞り込むのはとても大変である．したがって，検索時間も長くなってしまう．例として，2019 年の合計特殊出生率を調べるときに,「合計特殊出生率」をキーワード検索すると,「人口動態調査」がヒットする．この調査は，51,000 件の統計表が掲載されており，検索オプションを「データベース，ファイル内を検索」にして，キーワード「合計特殊出生率」で検索すると，約 26,000 件に絞り込まれる．そこから調査年を 2019 年にすると，約 1,000 件に絞り込まれる．これをさらに絞り込むためには，適切なキーワードを選択するなど, 経験から得られるノウハウが必要になる. e-Stat の長所である膨大なデータ量は, 逆に, 統計表の絞り込みの困難さを招いている.

(2) 専門用語の難しさ

e-Stat では，前述したように，事項名や表頭項目名，表側項目名などの専門用語を知らないと，検索に支障をきたすことが多い．さらに，日常ではあまり使用しない用語が出現することも多い．例えば，e-Stat でよく使用される「従業上の地位」という用語がある.「従業上の地位」とは，仕事をしている人が職場においてどのような地位であるかを区分したものである. 区分の例としては, 就業者や個人事業主, アルバイトなどがあるが，調査によって区分の範囲が違っていることがあり，どのような違いがあるか調べる必要がある.

関西学院高等部数理科学部[4]は，e-Stat が子供にとって使用するのが難しいので，小中学生のための統計情報ポータルサイト「e-Stat Junior」を提案している．しかし，筆者らの経験では，子供でなくても操作方法に慣れていない初心者にとっては操作するのが難しいと考える.

3.2. 検索の正確性の観点からの課題

検索の正確性の観点からは，検索漏れと検索ノイズの課題が存在する．検索漏れは，キーワード検索した場合，本来ヒットしなければならない統計表がヒットしない事象である．一方，検索ノイズは，本来ヒットしないはずの統計表がヒットする事象である．本研究では，特に，検索漏れについて考えることにする．検索漏れの起こる要因のひとつは，事項名と項目名の階層構造の不統一が挙げられる．もうひとつの要因は，事項名と項目名の表記不統一である．以下，具体的な例で説明する.

(1) 事項名と項目名の階層構造の不統一による検索漏れの具体例

キーワード「性別」で検索した場合，事項名が「男女」で項目名が「男」「女」の統計表はヒットしない．なぜなら，事項名にも項目名にも「性別」というワードがないからである．一方，事項名が「性別」で項目名が「男」「女」の統計表はヒットする．このように，事項名の表記の違いだけでヒットしたりしなかったりするのは好ましくなく，ともにヒットすべきである．したがって，この事象は，検索漏れが生じているといえる．また，図 4 のように，事項名が「職種_128 職種区分」で項目名が「プログラマー（男）」「プログラマー（女）」の統計表もヒットしない．これも性別に関する統計表なのでヒットすべきであり，検索漏れを起こしていると考えるべきである.

(2) 事項名と項目名の表記不統一による検索漏れの具体例

キーワード「性別」で検索した場合，事項名が「性」で項目名が「男」「女」の統計表はヒットしない．これは,「性別」と「性」の表記不統一による検索漏れである.

4. 性別に関する事項名と項目名の調査

本章では，「性別」に関する検索漏れの課題について調査した結果を述べる．調査の範囲は，「性別」に関するデータの多かった 5 省（総務省，経済産業省，文部科学省，農林水産省，厚生労働省）の基幹統計である．調査対象の統計調査は 41 件であり，事項名は 25,854 個，項目名は 1,559,628 個であった．

4.1. 事項名と項目名の階層構造の調査

「性別」に関する事項名と項目名の階層構造を調査した．事項名に性別に関する属性のみが入っており，項目名に性別の下位概念である「男」「女」などが入っているパターンは約 20 種類あることが判明した．その例の一部を図 7 に示す．①と②のように統計調査毎に事項名の表記方法が「性別」と「男女別」のように異なっている場合や，②と③のように同じ統計調査でも事項名が「男女別」と「男女」のように異なっている場合がある．⑤では，性別と学歴の 2 つの属性が事項名と項目名に含まれている．⑥では，項目名を見ると職種と性別の属性が含まれているが，上位概念である事項名には性別についての属性がない．このように，事項名と項目名の階層構造や名称が統一されていないことが判明した．

①学校基本調査（文部科学省）

性別	男
	女

②人口推計（総務省）

男女別	男
	女

③人口推計（総務省）

男女	男
	女

④科学技術研究調査（総務省）

男女別	男性
	女性

⑤賃金構造基本統計調査（厚生労働省）

性別_学歴	男（大学・大学院卒）
	女（大学・大学院卒）

⑥賃金構造基本統計調査（厚生労働省）

職種_128職種区分	プログラマー（男）
	プログラマー（女）

図 7　「性別」に関する事項名と項目名の階層構造

4.2. 同義語の調査

性別に関する同義語として，「性別」「男女」「男」「女」の 4 種類の単語の同義語があることがわかった．「性別」の同義語は上位概念である事項名で集計を行い，「男」と「女」の同義語は下位概念である項目名で集計を行った．その集計結果を表 1 と表 2 に示す．「性別」には 4 種類の同義語があり，「性」は 18 件とわずかしか出現しなかった．「性別」は 123 件，「男女」は 142 件，「男女別」は 80 件であり，「性」より多く使われている．「男」と「女」には，それぞれ 3 種類の同義語があり，「男」と「女」が 14,627 件で圧倒的に多く使われている．

表 1　「性別」の同義語

「性別」の同義語	事項名に含まれる件数	事項名の例
性別（代表語）	123	性別_学歴
性	18	性・年齢階級_001
男女	14	男女月３９150059
男女別	80	世帯主の男女別2010

表 2　「男」「女」の同義語

「男」の同義語	「女」の同義語	項目名に含まれる件数	事項名の例
男（代表語）	女（代表語）	14,627	学生数(計)【男】
男性	女性	1,376	従業者数_常勤_男性
男子	女子	161	事業従事者数_従事構成員_男子

4.3. 区切り符号の調査

事項名や項目名には，括弧やハイフンなどの区切り符号を含んでいるものが多い．これは，「プログラマー（男）」のように，職種と性別など，事項名や項目名に複数の属性を含む名称を表したいときに使用されている．そこで，どのような区切り符号が用いられているのかを調査した．その結果を表 3 に示す．例えば，事項名「性・年齢階級_001」は，「・（中黒）」によって性別と年齢を区別している．

表 3　区切り符号の例

（　）	全角	―	全角
(　)	半角	-	半角
【　】	全角	＿	全角
［　］	全角	_	半角
〈　〉	全角	・	全角
｛　｝	全角	□	全角空白

調査によって，以下の 3 種類の区切り方があることがわかった．

① 「他の事項名あるいは属性」を前に出し，括弧で「性別」を囲む方法（例：人口（男））

② 「性別」を前に出し，括弧で「他の事項名あるいは属性」を囲む方法（例：男（人口））

③ ハイフンなどの符号で「他の事項名あるいは属性」と「性別」を繋げる方法（例：人口-男）

調査の結果，①は約 8,500 件，②は約 500 件，③は約 2,000 件で，①と③がとても多く，②はわずかであった．

次に，「性別」に関する同義語の配置場所について調査した．配置場所とは，性別に関する同義語である「男」や「女」が事項名と項目名のどの場所に配置されているかということである．配置場所は，前，中，後の 3 種類である．4.1 節の同義語の調査において，項目名で一番多かった「男」と「女」を対象に，区切り方別に配置場所を調査した．その結果を表 4 に示す．単語を囲む方法は「中」と「後」が多いのに対して，区切り符号で区別する方法では圧倒的に「後」が多いことが判明した．

表 4　配置場所の調査結果

単語	区切り方	前	中	後
「男」 「女」	①括弧	10	1, 722	2, 026
	③符号	986	1, 381	8, 502

5. 解決策

本章では，検索漏れの解決策を提案する．解決策として，事項名に対する対策と，検索キーワードに対する対策の 2 つを組み合わせた方法を考える．

(1) 事項名に対する対策

事項名と項目名の階層構造の不統一や表記の不統一の問題を解決するために，項目名に「性別」の同義語が存在する場合，「性別」を事項名に追加し，それを用いて検索することにより，検索漏れを防ぐことが可能になる．そのイメージ図を図 8 に示す．従来の事項名①の項目名に「医師（男）」という「性別」の同義語があるため，事項名「職種」に対し事項名として「性別」を追加し，検索させるようにする．追加した事項名を拡張事項名と呼ぶことにする．

(2) 検索キーワードに対する対策

ユーザの検索キーワードに「性別」に関するワードが含まれていれば，全て「性別」で検索を行うようにする．例えば，「男」「男性」「男子」「男女」などが検索キーワードとして入力されたら，全て代表語の「性別」に変換し検索を行う．

従来の事項名で検索した場合と提案する拡張事項名で検索した場合でのヒット件数の調査を行った．事項名の 25,854 個に対して，従来の事項名と提案する拡張事項名において「性別」で検索した場合の結果を表 5 に示す．従来の事項名で検索をした場合は，212 件のヒット数であるのに対して，提案する拡張事項名では，項目名に性別に関する情報があれば事項名に性別を追加しているため，1,496 件のヒットとなった．すなわち，提案する拡張事項名で検索を行うことで約 1,300 件の検索漏れを防ぐことが可能になると考えられる．この予備調査より，本提案方式が有効であることが期待できる．

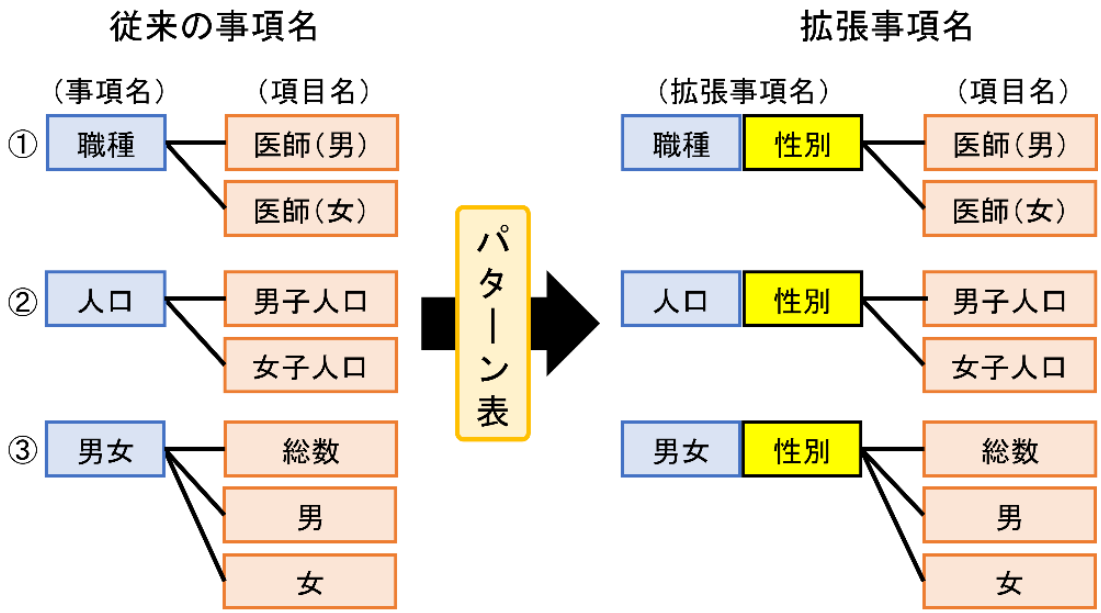

図 8　拡張事項名の追加

表 5　検索ヒット数

事項名	従来方式	提案方式
25, 854	212	1, 496

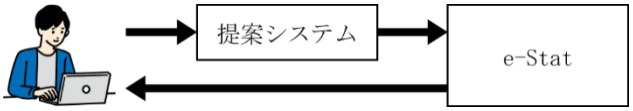

図 9　提案システムの利用イメージ

提案システムの利用イメージを図 9 に示す．提案システムでは，ユーザが入力した検索キーワードを同義語の代表語に変換し，提案システム内にある拡張事項名を用いて e-Stat アクセスし，ユーザの所望する統計表を得るという流れである．

6. まとめ

本研究では，特に「性別」に関して，e-Stat における検索漏れの原因を調査し，「性別」に関する事項名と項目名の階層構造や表記の違いを明らかにした．そして，その解決策についても提案した．今後，その解決策を実現するシステムを作成し，検証する予定である．また，この方法は，「性別」以外にも適用できると考えている．

参考文献

[1] デジタル庁, “オープンデータ，オープンデータ基本指針,” https://www.digital.go.jp/resources/open_data/, 2022.7.15 参照.
[2] 総務省統計局, “e-Stat 政府統計の総合窓口,” https://www.e-stat.go.jp, 2022.4.12 参照.
[3] 日本経済新聞, “政府統計，8 割がデータ検索できず　縦割りが浮き彫り,” 2021.9.1, https://www.nikkei.com/article/DGXZQOUA31AJD0R30C21A8000000/, 2022.7.6 参照.
[4] 関西学院高等部数理科学部, “小中学生のための統計情報ポータルサイト「e-Stat Junior」の提案,” STAT DASH グランプリ, 2016, https://www.e-stat.go.jp/api/sites/default/files/uploads/Policy-3_Sasaki_DAIJIN-1.pdf, 2022.6.1 参照.