「大学基本情報」の BI ツール利用の可能性と課題

― 複合的統計表のデータベース化とデータ解析に向けて ―

船守美穂（国立情報学研究所）
中鉢直宏（帝京大学）

１.本発表の目的と課題

大学改革支援・学位授与機構（NIAD）のウェブサイトには、国内国公立大学の学生数や教職員数などに関わるデータが「大学基本情報」として公開されている。これは、国の指定統計である「学校基本調査」の個票データを各大学から得ることで、これまでの集計値レベルの公表から一歩踏み込み、大学別／学部・研究科別／学科・専攻別にデータを公開したもので、各大学が自身の大学経営戦略を検討することを可能とした、極めて意義深いデータセットである。

一方、このデータセットは、「学校基本調査」という伝統的な統計表をベースとして作成され、利用の方法としても所謂統計表として、人間の読みやすさや理解のしやすさに配慮して作成されたものである。このため、統計表の行や列の途中に「小計」の欄があり、また見出し行も複合的かつ複数行にわたっているなど極めて複雑で、これらのデータをデータベースのように BI ツールなどで機械的に読み込むことが困難なレイアウトとなっている。また単一の年度において 20 以上の表が提供されているが、それらの表の間のデータ項目の用語統一も十分になされていないため、複数の表を機械的に相互連携させることも現状では出来ない。さらに表が年度ごとに公表されているため、単一の表を複数年にわたって解析することも出来ない。しかし一方でこれらの表は、学科・専攻別のデータが列挙され、行数が膨大で、表形式のままで人間が表の内容を読み取るには限界がある。

このため、この「大学基本情報」のデータセットを可視化とデータ解析が可能となるように、BI ツールで読み込むためのデータ加工を行った。この加工したデータセットは公開し、全国の大学 IR 担当者等の利用に供する予定である。本小論は第一義的には、このデータセットを利用する担当者が、どのようなデータ加工が行われたかの把握可能なように、データ加工のプロセスやそのような加工の判断理由について、ドキュメント化したものである。また今年度以降、新たなデータセットが加わった際も、このドキュメントをもとに同様の加工ができることを期待している。さらに、本「大学基本情報」の BI ツール利用に向けての加工の手法が、一般的な複合的統計表のためのデータベース化とデータ解析においても汎用的に利用可能であることを、期待している。

２.「大学基本情報」データセットの概要

2016 年度末において、表 1 に挙げるデータセットが 2012－16 年度の 5 年度分、公開されている。一方、年度によっては欠落しているデータセットや、集計方法が途中から変更となったデータセットがある。集計方法が変更となったデータセットについては、他のデータセットから類似のデータセットを生成可能な場合、そのようなデータを挿入した。

表１：「大学基本情報」の年度別データセットと、リシェイプ後のデータセット

（ファイル名）	2012	2013	2014	2015	2016	リシェイプデータ
(7-A) 学生数	○	○	○	○	○	07go1_学生数
(7-B) 教員数（本務者）	○	○	○	○	○	07go2_教員数（本務者）組織区分別
(7-1) 教員数（本務者）（再掲）	○	○	○	○	○	07go3_教員数（本務者）教員状態別
(7-Z) 教員数（兼務者）	○	○	○	○	○	07go4_1_教員数（兼務者）職位別
						07go4_2_外国人教員数（兼務者）
(7-C) 職員数	○	○	○	○	○	07go5_1_職系別職員数
						07go5_2_医療系職員数
(8-D) 学科別学生数	○	○	○	○	○	08go1_1_学科別学部生数
	○	○	○	○	○	08go1_2_学科別学部入学者数
	(8-E) 入学状況から抽出（学科情報なし）			未公開	○	08go1_3_学科別学部入学志願者数
(8-2) 学科別学生数のうち休学者数	○	○	○	○	○	08go2_学部別休学者数
(8-3) 学科別学生数のうち最低在学年限超過学生数（編入学者は除く。）	○	○	○	○	○	08go3_学部別最低在学年限超過学生数
~~(8-E) 入学状況~~	○	○	○	未公開	×	~~08go4_学部入学状況（志願者数、入学者数）~~
(8-O) 年齢別入学者数	×	×	×	×	○	08go4_1_年齢別学部入学者数
	(8-E) 入学状況から抽出			未公開	○	08go4_2_出身高校種類別学部入学者数
	×	×	×	×	○	08go4_3_学部留学生入学者数
(8-G) 出身高校の所在地県別入学者数	○	○	○	○	○	08go5_出身高校の所在地県別学部入学者数
(8-7) 専攻科・別科及び聴講生等の学生数	○	○	○	○	○	08go6_専攻科、別科、学部科目等履修生等の学生数
(8-R) 短期大学・高等専門学校・専修学校（専門課程）からの編入学者数	○	○	○	○	○	08go7_学部編入学者数
(9-H) 専攻別学生数	○	○	○	○	○	09go1_1_専攻別院生数
						09go1_2_専攻別社会人院生数
(9-4) 専攻別学生数のうち休学者数	○	○	○	○	○	09go2_研究科別休学院生数
(9-5) 専攻別学生数のうち最低在学年限超過学生数（編入学者は除く）	○	○	○	○	○	09go3_研究科別最低在学年限超過院生数
(9-I) 入学状況	○	○	○	○	○	09go4_1_専攻別大学院入学者数
						09go4_2_専攻別大学院入学志願者数
(9-S) 年齢別入学者数	○	○	○	○	○	09go5_1_年齢別大学院入学者数
						09go5_2_大学院社会人入学者数
						09go5_3_大学院留学生入学者数
(9-8) 科目等履修生等の学生数	○	○	○	○	○	09go6_大学院科目等履修生等の学生数
(10-J) 学科別学生数、入学状況（本科）	○	○	○	未公開	○	10go1_1_学科別本科学生数
						10go1_2_学科別本科入学者数
						10go1_3_学科別本科入学志願者数
(10-6) 学科別学生数のうち休学者数	○	○	○	未公開	○	10go2_本科休学者数
(10-K) 出身高校の所在地県別入学者数	○	○	○	未公開	○	10go3_出身高校の所在地県別本科入学者数
(10-9) 専攻科、別科及び科目等履修生等の学生数	○	○	○	未公開	○	10go4_専攻科、別科、本科科目等履修生等の学生数
(10-Q) 年齢別入学者数(再掲)	×	×	×	×	○	10go5_1_年齢別本科入学者数
	(10-J) 学科別入学状況から抽出			未公開	○	10go5_2_出身高校種類別本科入学者数
	×	×	×	×	○	10go5_3_本科留学生入学者数
(11) 国費留学生、私費留学生、留学生以外の外国人学生	○	○	○	○	○	11go1_外国人学生数
(11 別掲) 国費留学生、私費留学生、留学生以外の外国人学生（専攻科・別科の学生、科目等履修生・聴講生・研究生	○	○	○	○	○	11go2_外国人学生数（専攻科、別科、科目等履修生等）
(20) 学校施設	○	○	○	○	○	20go_1_学校土地の用途別面積（㎡）
						20go_2_学校建物の用途別延面積（㎡）
						20go_3_厚生補導施設の延面積（㎡）
						20go_4_学校建物の構造別延面積（㎡）
						20go_5_学校建物増減面積（㎡）
						20go_6_職員宿舎土地の用途別面積（㎡）
						20go_7_職員宿舎建物の用途別面積（㎡）
(30-1) 状況別卒業者数、入学年度別卒業者数	○	○	○	○	○	30go1_1_状況別卒業者数
						30go1_2_就職状況別卒業者数
						30go1_3_博士課程満期退学者数
						30go1_4_ポストドクター数
						30go1_5_入学年度別卒業者数
(30-2) 職業別就職者数、産業別就職者数	○	○	○	○	○	30go2_1_職業別就職者数
						30go2_2_産業別就職者数

３.「大学基本情報」データセットの BI ツール利用にあたっての課題と対応策

NIAD のウェブページで公開されている「大学基本情報」は冒頭に説明したように、人が見ることを想定とした統計表として基本的には作成されており、BI ツールの利用に適していない。また、複数表を連携して分析する上でも不整合があり、データベースとしての体をなしていない。

以下に、「大学基本情報」のデータセットを BI ツールで利用可能とするために必要な加工の観点を挙げる。

A. BI ツールで利用可能とするためのデータ加工

A-1. データセットの年度別欠落や集計方法の変化への対処

年度によっては欠落しているデータセットや、集計方法が途中から変更となったデータセットがある。これらについては、過去のデータセットを可能な限り最新のデータセット(2016 年度)のフォーマットに合わせて、整形した。しかし、データが存在せず、そのような加工が不可能であった場合もある（表１「×」)。

今後、大学改革支援・学位授与機構によりデータセットが毎年度追加されていくにあたり、集計方法や公開されるデータセットが変更となってくると、今回作成したデータセットを全て作り替える、あるいは追加されたデータセットを現在のフォーマットに変換する必要があることには、留意が必要である。この場合、データセットの BI ツール利用への加工の手間が煩雑な上、そのようなフォーマットの統一化が不可能なデータセットが複数出現することは、避けられない。

A-2. データセット間の用語（列見出し）の不統一への対応

BI ツールを利用する魅力の一つに、複数の表を連結させてデータ解析できることが挙げられる。たとえば、学科別 ST 比などは、学生数と教員数の表を連結させることにより、算出、表示可能となる。他方、これを実現するためには、異なる表の見出し、および、その値の表記が同一でないと、これらの表を機械的に連結することができない。学科別 ST 比の例では、「学科･専攻」という見出しと各データが、教員数と学生数の表の双方で、統一されている必要がある。

表の見出しは、もともとは個々の表において理解しやすいように振られており（特に、見出しが長くなりすぎないように、簡略に表現されている)、複数表横断的に統一性をもって振られているわけではない。今回のデータセットの加工においては、「大学基本情報」の 20 以上の表横断的に統一感をもった列見出しを振った。なお、上述の例の「学科・専攻」の各データ（＝学科名）の表記ゆれには対処せず、学科名を用いた複数表の連結には、「学科・専攻コード」が利用されることを想定している。

A-3. 複数の異なる集計値を含む単表の分割

「大学基本情報」の表には、異なる性格の集計値が単一の表に含まれているものがある。たとえば、「入学状況」の表には、「入学者数」と「入学志願者数」のデータが含まれている。このように、異なる性格の集計値が単一の表に含まれる場合は、表を複数に分割することとした。表１の「リシェイプデータ」に、複数に分割した表を一覧した。

A-4. 複合見出しの単行化

「大学基本情報」の表の一部は、人間の読みやすさに配慮して、列見出しが複数の行に

わたっている。たとえば学生数の表において「年次」や「性別」は、学生数の列の見出しに組み込まれており、複数年次の男女の学生数が横比較可能となっている。

一方 BI ツールを利用するためには、集計値を示す列は 1 列にまとまり、残りの表は属性を示す列として構成されている必要がある。つまり「年次」や「性別」などの属性は、列見出しとして組み込まれているのではなく、列情報として各行に、提示されている必要がある。

なお NIAD で公表する「大学基本情報」の表には、隠れた見出しの行が存在し、複数行にわたる列見出しを単一の見出しで示すラベル（たとえば「1 年_男」「3 年_女」等）が提供されているが、これは利用せず、見出しにおける全ての属性情報を列情報として組み込む作業を行った。また、「集計対象」という列を新たに設け、具体的に何を集計しているのか（たとえば「学生数」「休学者数」等）を記載した。

見出しにおける属性情報を列情報として組み込むにあたっては、“Tableau Excel Add-In”を利用した[1]。このツールは整合性のとれたクロス集計を BI ツールで扱うデータ形式に、ワンタッチで並び替えることができる。今回の作業は、このツールを利用することにより、大幅に軽減された。

B. BI ツールの利用しやすさのためのデータ加工

B-1. 分かりやすく、規則性あるファイル名の付与

NIAD で公表されている「大学基本情報」の各表のファイル名は、年度とコードとが組み合わさったものとなっている。またそのコードは、データの元となった「学校基本調査」に準拠しており、(7-A), (7-B), (7-1), (7-Z), (7-C)などという並びとなっている。これは「学校基本調査」を知らない者にとっては意味が理解できないだけでなく、パソコン等でファイルが表示される場合、ファイルの並び順がおかしくなる。また、A-4 節に示したように、今回は一つの表を複数に分割したこともあり、ファイルが順番に表示されるようにコードを振り直した。またファイル名からファイルの内容がくみ取れるよう、ファイルの保持するデータの内容についても、ファイル名に含めた。

B-2. コードの分割と内容情報の付加

たとえば「研究科番号」のコードなど、1 桁目が課程（修士課程、博士課程等）、下 3 桁が分野を擬似的に意味すると把握されたものについては、コードを 2 つに分けた。また、人の可読性に配慮して、そのコードの内容を表現する列を付加した。ものによっては、コード情報と文字情報を組み合わせ、文字情報の並びが分かりやすくなるよう工夫をした。

「所在地コード」や「課程別」、「学校種別」などのコードの内容については、「学校基本調査の手引き」を参照した[2]。他方、情報として最も有用であるはずの学部・研究科や学科・専攻の分野を示すコードについては、コードと内容の対応表が公開されていない。文部科学省生涯政策局政策課調査統計企画室学校基本調査係によると、コード表は公開はしていないものの、「学科系統分類表」（コードなし）は公開しており、これを参考にされたいとのことであった[3]。今回のデータ加工では、「大学基本情報」に元からあった、学部・研究科や学科・専攻の分野を示すコードと、各大学において手入力された学部・研究科や学科・専攻の名称（表記ゆれ含む）との双方を残し、これらの対応関係が確認できるようになっている。

B-3 小分類コードへの、大中分類の付加

たとえば「所在地（都道府県＋政令市)」や「国籍」などは区分数が多く、BI ツールの利用にあたって人による選択が困難となるため、大中分類を付加した。「所在地（都道府県＋政令市)」については「所在地県」や「八地方区分」、「国籍」については「地域区分」などが付加され、絞り込みが容易になっている。

B-4. 合計行および無回答行の削除

「大学基本情報」の表には、人の可読性に配慮してときどき、小計や合計などの行が挿入されている。これらは「小計」等のフラグを立てておけば、その他の行と分離可能であるが、BI ツールでデータを分析する際に意図せず、このような行も含めて合算等をしてしまう危険性もあるため、それら行については削除をした。

また無回答行は、NIAD で公表する「大学基本情報」の表上「０（ゼロ)」が入力され、「０（ゼロ)」と回答されたものと混在しているケースが見られる。このことにより BI ツールを利用した分析において不整合が起きる可能性があるため、もともと「無回答」であったと把握可能な行については、削除した。

４．BI ツール向けに変換後の「大学基本情報」データセット利用にあたっての留意点

（１）データセットの年度別欠落や集計方法の変化

A-1 節に詳説したように、データセットにより年度別の欠落や集計方法の変化があり、これらについては可能な限り対処はしたものの、データを何も考えずに利用や可視化すると、データの飛び等で問題が生じる可能性がある。

（２）大学別の集計―「番号別大学名」の利用の勧め

「大学名」のフィールドは、各大学が入力した値がはいっており、表記ゆれがあることがある。このため、元データから存在する「学校調査番号」と「大学名」とを機械的に組み合わせて作成した「番号別大学名」を、大学別の集計等では利用することを推奨する。

（３）「学部・研究科」「学科・専攻」の分野別分析における留意点

「大学基本情報」のデータセットの魅力の一つは、「学部・研究科」や「学科・専攻」別の個票データが存在し、これらの分野別にデータ分析が可能なことである。同じランクの大学とみなされていても、学部・研究科構成等が異なる大学は、学生数や教職員数などの基本的に数値さえも、相互比較することができない。

一方「学部・研究科」や「学科・専攻」の名称は、たとえば同じ経済学系の学科が、大学によって「経済学」や「経済学類」とされるなど、同じ内容であっても名称が異なり、学科名称でグルーピングすることができない。しかし学科・専攻コードではこれらは“C203”で統一されているため、グルーピング可能である。他方、コード側も極めて細かく振られており、たとえば経済・経営系の学科では、「商学（C201)」「経済学（C203)」「経営学（C205)」「経営経済学（C208)」「国際経済学（C209)」「応用経済学（C215)」「経営情報(C217)」「経済工学(C219)」「国際商学(C224)」「国際経営学(C225)」「流通学(C226)」「経済情報学（C227)」などとなっており、単一のコードに依存するのは危険である。分析したい対象に応じて、たとえばコードの１桁目のみを参照する、「経営」という KW でグルーピングする等の配慮が必要である。

（４）空欄データへの注意

BI ツールを利用すると、データの入っている表を確認せずに、グラフ表示のためのデータラベル（見出し）のみを操作する危険性が高い。しかしたとえば「教員数」のデータの大部分は学科・専攻別に存在するものの、学部・研究科レベルで教員の所属が管理されていたり、大学執行部など大学本部に所属があり、「学科・専攻」欄は空欄である大学も存在する。これらに気づかずに、教員数を学科・専攻別に集計すると、大学別の総教員数と齟齬が生じるなど、問題が発生する危険性がある。これを回避するには、頻繁にデータに立ち戻りながらデータの可視化を行うなどの注意が必要である。

５．「大学基本情報」データセット利用上の課題と今後の展開

「大学基本情報」のデータセットを BI ツールで利用可能なように整形したが、これを大学 IR に利用するという観点からみた場合、課題も多く残されている。

まず「大学基本情報」に含まれるデータは、学生数や教員数などの基本的なデータのみであり、たとえば大学の競争力などを示す指標は明示的には存在しない。このため、これらデータセットを単に可視化しても、つまらないだけである。しかしたとえば女子学生比率や留学生比率、休学者数、学年別の留年者数など、細かく分析すると大学の現状把握につながる指標は抽出可能であり、つまり、どのような目的でデータを分析するかというリサーチクエスチョンの立て方において、IRer の力量が大きく問われる状況となっている。

こうした個人の力量への依存を多少でも回避するためには、たとえば、ある程度規格化された「大学のヘルスチェックレポート」の作成・提示、大学の多様性や学生の標準年限内修了状況などの「大学の課題別テーマ」に基づいた一連の分析流れの提示などを明確にし、示していくことが必要と考えられる。

同時に、4.(3)節で示した、「学部・研究科」や「学科・専攻」の分野別分析についても、個々の大学や担当者の判断でグルーピングするだけでなく、ある程度大ぐくりであっても、よく使われる可能性のあるグルーピングは事前に提示することが重要である。

なお今回は「大学基本情報」という、国の指定統計である「学校基本調査」をベースとした一連の統計表を BI ツール利用可能なようにデータ加工したが、ここで見いだされた 3 節に挙げたデータ加工の手続きは、他の伝統的な統計表のデータ加工にも適用できる方法である。近年、「公的統計調査の調査票情報等の学術研究等への活用」に見られるように、伝統的な統計調査の個票が、一定程度の匿名化措置はなされていても、利用可能となっていく上で、有用な知見となっていくと考えられる。

【参考文献】

[1] Tableau Community, “Tableau Add-In for Reshaping Data in Excel”
（http://kb.tableau.com/articles/knowledgebase/addin-reshaping-data-excel）
[2] 文部科学省「平成 28 年度 学校基本調査の手引（大学，短期大学，高等専門学校）」
（http://www.mext.go.jp/b_menu/toukei/chousa01/kihon/sonota/1355787.htm）
[3] 文部科学省「学科系統分類表（高等教育機関）」
（http://www.mext.go.jp/b_menu/toukei/chousa01/kihon/shiryo/sh_detail/1375044.htm）