大規模データ群のデータベース化について
ー河川水辺の国勢調査の過去データの真正化ー

Building a large-scale database
— Validation of data collected through National Census on River Environments —

研究第四部　研　究　員　**松間　充**
企画・広報部　副　参　事　**大石　三之**
審　議　役　**小川　鶴蔵**

河川水辺の国勢調査（生物調査）ではより効率的なデータ活用のために、平成12年度調査以降は調査結果の電子化が図られている。しかし、平成11年度以前の調査データ（以下、過去データ）についてはデータベースに搭載可能なデータ形式にはなっていないため、データの変換等が必要である。本研究では、河川水辺の国勢調査の過去データの「真正化」を例に、膨大なデータ群をデータベース化する際の手法について検討することを目的とした。

過去データの真正化においてはデータ入力時のエラーの解消に多くの労力が費やされた。このような全国規模のデータベースを作成する際には、単純なエラーが出ないように対策を講じることが重要であることが示唆された。

キーワード：河川水辺の国勢調査、過去データ、真正化、河川環境情報システム、データベース、水情報国土データ管理センター

Data collected through the National Census on River Environments (biological surveys) have been turned into electronic form since the 2000 survey in order to make more efficient use of data. Since, however, the 1999 and earlier survey data (”past data”) are not in a format compatible with the database, they need to be subjected to data processing including data conversion. The purpose of this study is to study methods for building a database from a large number of data sets, focusing on the validation of past data obtained from the National Census on River Environments.

In validating the past data, considerable effort was spent in eliminating data entry errors. This suggests that when building a database on a nationwide scale, it is important to take measures to minimize simple errors.

Key words : National Census on River Environments, past data, validation, River Environmens Information Systems, database, National Land with Water Information Data Management Center

1．はじめに

平成２年度から開始された「河川水辺の国勢調査」は、河川に生息・生育する動植物の相の把握という点で大きな成果を挙げている。しかし、その成果は毎年「年鑑」という形で発行されてきたものの、この「年鑑」は、

・公開しているデータの項目数の割合が全取得データに比べ４割程度である
・調査実施当該年度のみのデータ発行であり、過去の調査データも含めた対象河川のデータを一括して扱うことができない

などの理由により十分活用されていない状況にある。

このような背景から平成12年度より、調査データの公開システムの運用に向けたデータの電子化が図られ、現在では、全国にある国土交通省の河川事務所（河川管理者）においては河川環境情報システムにより、また一般向けには国土交通省の「水情報国土データ管理センター」からのデータ提供により、データの検索や時系列的な分析等によるデータ利用が容易となっている。

平成11年度以前の過去データについては、データ取得の際にシステム運用を意識した電子化が図られていないことから、平成12年度以降の調査データと同様にシステム運用できる形式のデータを作成する必要がある。よって、このシステム運用が可能となる形式のデータを元の調査データから生成する「真正化」作業を行うものとする。なお、真正化とは、調査データが最新の生物種目録に基づいており、且つ、最新版である平成９年度調査マニュアルに沿ったデータ構造でデータベース化、GIS 化され、正規のデータとして河川環境情報システムでの運用ができるようにすることである。

上記のような背景から、本研究では、河川水辺の国勢調査の過去データの真正化検討を例にとり、膨大なデータ群をデータベース化する際の手法について検討することを目的とする。

2．河川水辺の国勢調査におけるデータの真正化

2－1　対象過去データ

河川水辺の国勢調査は、平成２年度から実施されているが、平成12年度以降の調査データは既にシステム運用に向けた電子取得に移行していることから、今回の検討では平成２～11年度の調査データが真正化の対象となる。

ただし、平成４年度以前の調査データについては、調査マニュアルが未だ十分熟成していなかったこと等の理由から平成５年度以降の調査データを対象とした。

表－１に全国の一級河川で平成５年度以降に実施された河川水辺の国勢調査データを、地方毎、調査項目毎に真正化対象となる過去データの数として整理した。全国の対象データ数の合計は1725であった。

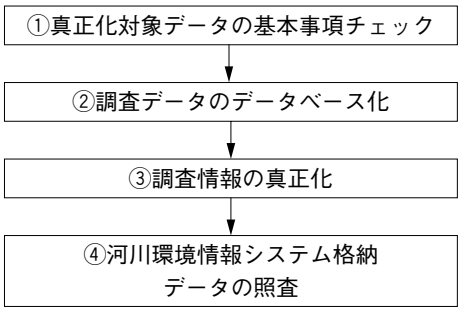
表－１　真正化対象過去データ数

調査項目／地方	魚介類	底生動物	植物	鳥類	両・爬・哺	陸上昆虫類等	合計
北海道	39	30	27	30	26	25	177
東北	42	28	32	34	32	32	200
関東	39	32	37	36	29	28	201
北陸	36	26	27	27	28	27	171
中部	48	33	32	29	35	40	217
近畿	65	38	34	33	34	34	238
中国	38	30	26	27	23	26	170
四国	16	16	15	14	14	14	89
九州	61	37	36	53	37	38	262
合計	384	270	266	283	258	264	1725

※データ数：１水系１調査項目１調査を１データとして計上

2－2　過去データの真正化手順の検討

過去データの真正化の手順は、図－１のとおりであり、「河川水辺の国勢調査システム過去データ入力マニュアル」（以下、過去データ入力マニュアル）で電子化された調査データをシステム運用できるデータ構造に置き換えるためには、図－１に示す４つの手順が必要である。



図－１　過去データの真正化フロー

それぞれについて作業の概要を以下に示す。

（１）真正化対象データの基本事項チェック

過去データ入力マニュアルにより電子化された真正

化対象データのデータベース化にあたって必要となる、電子ファイルの有無、調査回数・調査地点数等の基本情報についてファイル間整合の確認を行う。

（2）調査データのデータベース化

「河川環境情報システム」のデータベースに格納できるようなフォーマットにデータ変換を行い、河川環境情報システムに格納する。

（3）調査情報の真正化

データベースに格納されたデータのうち、生物名に関する精査を行う。精査には「河川水辺の国勢調査のための生物リスト」（以下、生物リスト）の最新版である平成12年度版を用い、リストと過去データを照合させることにより、整合がとれなかった種名について変更を行う。変更結果については変更履歴を作成し、原記載との関係を明らかにしておく。

（4）河川環境情報システム格納データの照査

「過去データ入力マニュアル」に基づき作成された真正化前のデータと河川環境情報システム格納データ（真正化後のデータ）に不整合がないことを確認するために照査を行う。

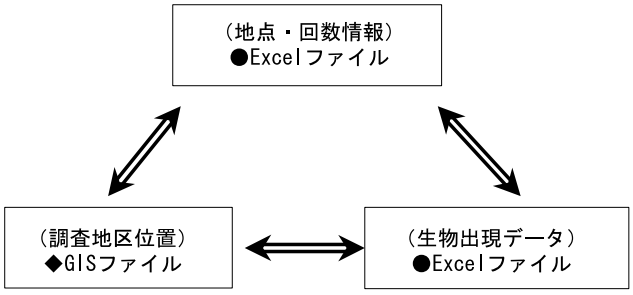
2−3　真正化作業

対象データは、各調査の報告書（紙ベース）を過去データ入力マニュアルに基づいて電子化したものであるが、真正化作業それぞれの作業段階で問題となった代表的な事例を以下に示す。

（1）真正化対象データの基本事項チェック

真正化対象となるデータは、過去データ入力マニュアルに基づいて作成した種名や調査地等の文字と数値からなる「EXEL データ」、調査地点などを地図データとして示した「GIS データ」、そして「写真データ」の主に3種類である。これらのデータはファイルの有無をチェックし、またそれにあわせて調査回数・地点数といった調査の基本的な情報についてもそれぞれ整合がとれているかの確認を行った。

調査回数・地点数チェックの例について図−2に示す。

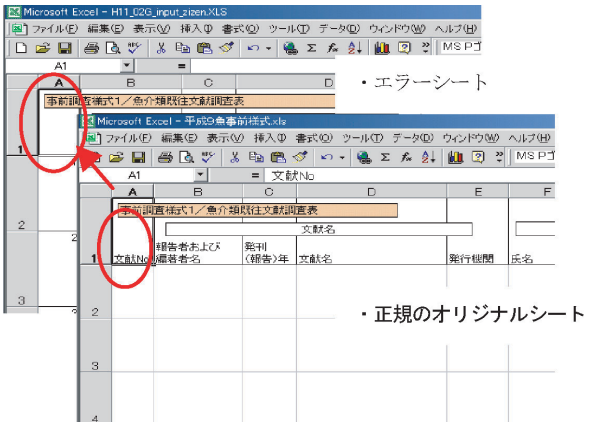


図−2　調査回数・地点数のチェック

データはいくつかのファイルに分かれて存在している。それぞれのファイルは調査回数等の情報により結びついているので、ファイル数に過不足があった場合には作業に支障がでる。ここでは、実際の調査回数に見合う数のファイルがあるかなど、基本的な情報について整合がとれているかどうかについてチェックを行った。整合がとれていないものについては問い合わせ確認を行い修正した。

（2）調査データのデータベース化

ここでは、データをデータベース形式に変換することを主な目的としているが、データ自体に何らかの問題がある場合は正常に変換することができない。データの変換時にファイルフォーマットの正当性、論理的な誤り、調査データの欠落についてチェックを行った。ファイルフォーマットのチェックについて、図−3に例を示す。



図−3　ファイルフォーマットチェックの例

過去データ入力マニュアルでは、決められたフォーマットに情報を入力するようになっている。データ入力者がそのフォーマットに変更を加えたため、データベースへ取り込む形式にうまく変換できないケースが見られた。図−3では、「文献 No.」という文字が削除されているために変換できない例を示している。このほかに行や列を挿入したり削除したりといった例が見られた。これらについては手作業により修正を行った。

また、現行の平成9年度版マニュアルでは、調査時の環境情報などは候補となる環境の母集団から選択する形式をとっており、河川環境情報システムではそれらをマスタとして管理している。例えば図−4の右表がマスタにあたり、調査者は調査時の環境に合わせてこのマスタの中から選んで記録することになる。しかし、過去データにおいてはマニュアルの違いから必ずしもマスタ通り記載されていないため、検討して現在のマスタにあわせるように統一化を行った。（図−4左）。

No.	報告書記載内容	統一登録内容（マスタ）
1	砂	地上
2	地	
3	土	
4	土・空	
5	畑	
6	田	
7	堤	
8	草	草上
9	木	樹上
10	不	
11	林	
12	木・空	

No.	報告書記載内容	統一登録内容（マスタ）
13	他・空	空中
14	空・他	
15	空・木	
16	土→空	
17	水	水面
18	ブロック他	その他
19	上	
20	電柱	
21	物	
22	○	不明
23	そ	
24	ほ	

鳥類位置マスタ
地上
草上
樹上
空中
水面
水中
その他
不明

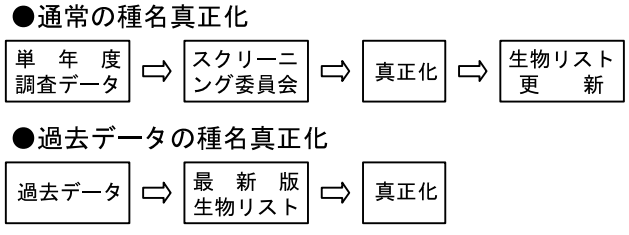
図－４　マスタによるデータ変換内容の例

データベースで入力必須項目となる項目が未入力であった場合もエラーとして認識された。例えば魚介類調査の場合調査値の水温や流速は必ずデータとして入れなければならないが、それらのデータが未入力であった場合にはデータベース形式に変換することができない。そのため、「99」など明らかにダミーとわかるようなデータを記入し、作業後の確認が行えるよう変更履歴を作成した。

このようにエラーのあるファイルがかなりの頻度で発生したため、データ入力者との確認作業が頻繁に行われた。また、これらのエラーが発生するたびにデータベースの形式に変換する過程で止まってしまい、その都度確認し手作業により修正を行わなければならないため、全作業の中でも相当な労力を必要とした。

（３）調査情報の真正化

河川水辺の国勢調査では、スクリーニング委員会という学識経験者等により構成された委員会により種名の真正化が行われる。今回の過去データの真正化においては、最新版の生物リストと照合することで真正化を行った。ここで、スクリーニングとは専門家による種名や分布の精査を意味する。図－５に通常の種名真正化の流れと今回の流れの比較を示す。



図－５　種名真正化の流れの比較

通常単年度の調査結果の真正化では、主に新規に確認された種の種名や分布に関してスクリーニング委員会にて精査を行う。そこで確認されることにより生物リストに新規確認種が追加され、更新される。生物リストは毎年の河川水辺の国勢調査の調査結果を受けているため、最新の生物リストはこれまでのスクリーニング結果を全て反映したものといえる。

今回は、過去データにおける確認種を現在の生物リストと照合し、一致しない種（目録未掲載種）について種名の修正等を行った。目録未掲載の理由は以下のものがあげられる。

①分類学の発展により調査時の種名に変更が生じた場合

②入力した種名に誤りがあった場合

③種名の表現が河川水辺の国勢調査のルールに従っていない場合

表－２に種名の修正例を示す。一番上の例は、平成11年に刊行された文献により、種名（学名）の変更が生じたので修正した例である。２番目の例は根拠となる資料が無いため、属という上位の階層にまとめた例である。３番目は同定の間違いではないが、河川水辺の国勢調査では種名の記述方法が統一されているので、そのように修正した例である。最後に、これが一番多くみられたが、種和名や学名の入力ミスを修正した例である。これには、種名の最後に半角スペースが含まれるなど、一見判別できないようなエラーも多数含まれる。このようなエラーの発見にはコンピュータによる自動チェックが有効であった。

データの修正を行った場合は、全て修正履歴を作成し、原記載との関係を確認できるようにした。

表－２　種名修正の例

元データの種名	真正化後の種名	変更理由
Agabus miyamotoi	Agabus optaus	甲虫図鑑Ⅱ　4刷（平成11年）により種名変更
イトウオオアリ	Camponotus 属の一種	出典不明につき上位分類に変更
Leersia sp.	Leersia 属の一種	不特定種の記述方法を統一
ネスミホソムギ	ネズミホソムギ	和名入力ミスを修正

２－４　河川環境情報システム格納データの照査

基本的には前作業までを行うことで河川環境情報システムでの運用は可能となるが、最後に真正化作業の前後でデータが正しく変換されているかどうかのチェックを行うこととした。データ照査方法については一般的な方法が確立されていないため、独自の方法を検

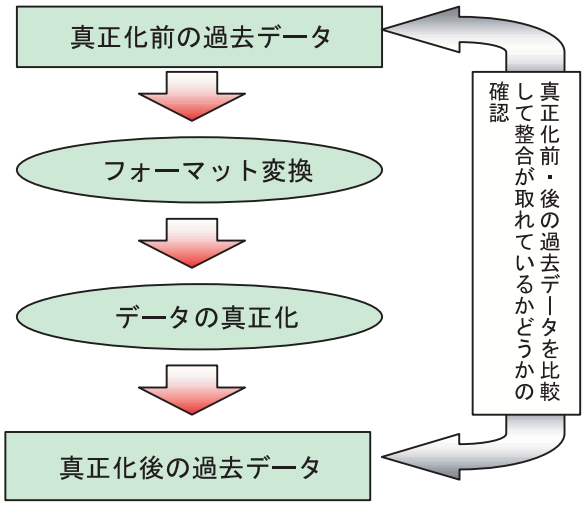
討した。ここではその内容について詳細に記述する。

（1）電子データに関する照査方法の現状

データ入力や変換といった作業の照査方法については、標準化した考え方が存在していないのが現状である。このためデータの照査は、現在 GIS 総プロで検討が進められている GIS データの照査方法の検討、研究結果を参考にして独自の方法について検討した。GIS 総プロとは GIS を活用した次世代情報基盤の活用推進に関する研究（国土地理院企画部測量指導課が中心となって推進中）である。

（2）過去データの照査の概要

作業は真正化前と真正化後の過去データを相互に比較することにより、適正にデータが変換され、整合がとれているかどうかのチェックを行うことである。図－6 にその概要を示す。

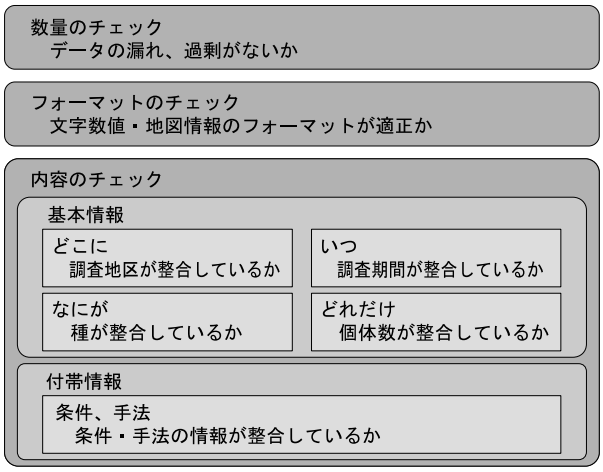


図－6　格納データの照査の概要

なお、前述の通りデータベース化の作業時にデータの修正・変更を行っているため、単純に比較すると真正化の作業前後でデータに違いが生じている場合がある。このような場合については、修正・変更履歴によって前後の関係の確認がとれれば、整合はとれているものと判断した。

（3）照査の視点

照査は、調査業務に関する基本事項である調査数量に関する事項、データフォーマットに関する事項、調査内容に関する事項、の大きく 3 つに分けて考えることができる。それぞれ、図－7 に示す視点でデータの照査を行うものとした。



図－7　照査の視点

（4）照査の方法

調査数量及びデータフォーマットについては、調査データの基本事項であることから全ての調査データを対象とした全数検査を行うこととした。また、調査地区、調査時期、種名、個体数、調査環境等、調査手法などのデータの内容に関する項目については、生物情報として基本情報であり、データ数が膨大な数であることから、その内容について整合を確認するため抜取検査を行うこととした。表－3 に照査方法の設定に関する検討結果をまとめた。

表－3　照査方法の設定

照査項目		照査方法	照査方法の設定理由
数量のチェック	調査回数、地区数	全数検査	調査データの数量は業務の基本事項であるため全数検査を行う。
フォーマットのチェック	文字数値・地図情報	全数検査	文字数値及び主題図のデータが遵守すべき関係やルールを満足しているかの基本的な検査であるため全数検査を行う。
内容のチェック	調査地区	抜取検査 ①1 調査業務毎に 1 サンプルを抽出 ②その上で、地方整備局内の全ロット数（各調査業務の地区数×回数の総計）に対する抽出サンプル数の比率が 5％を下回らないようにさらにサンプルを抽出	地区情報、時間情報、種の分類や名称、個体数、調査環境等や手法は、生物調査の基本情報であることから、その内容について整合を確認するため抜取検査を行う。
	調査時期		
	種名		
	個体数		
	調査環境等		
	調査手法		

（5）抜取検査における照査データの抽出方法

抜取検査の対象となる照査データの抽出は、以下の方針により行う。

・抜取検査は全調査データを生物調査項目毎に分けて行う。
・抜取検査の検査単位（サンプル）は、生物調査項目毎に分けられた全調査データに対して、“調査地区ごとに従属する調査回データ”を基本単位とする。
・１調査業務ごとに乱数表を用いて１サンプルを抽出する。
・その上で、地方整備局内の全ロット数（各調査業務の地区数×回数の総計）に対する抽出サンプル数の比率が５％を下回るようであれば、残りの全ロットを対象に乱数表を用いてサンプルを抽出し、全ロット数に対する抽出サンプル数の比率が５％以上となるようにする。

（6）照査の内容

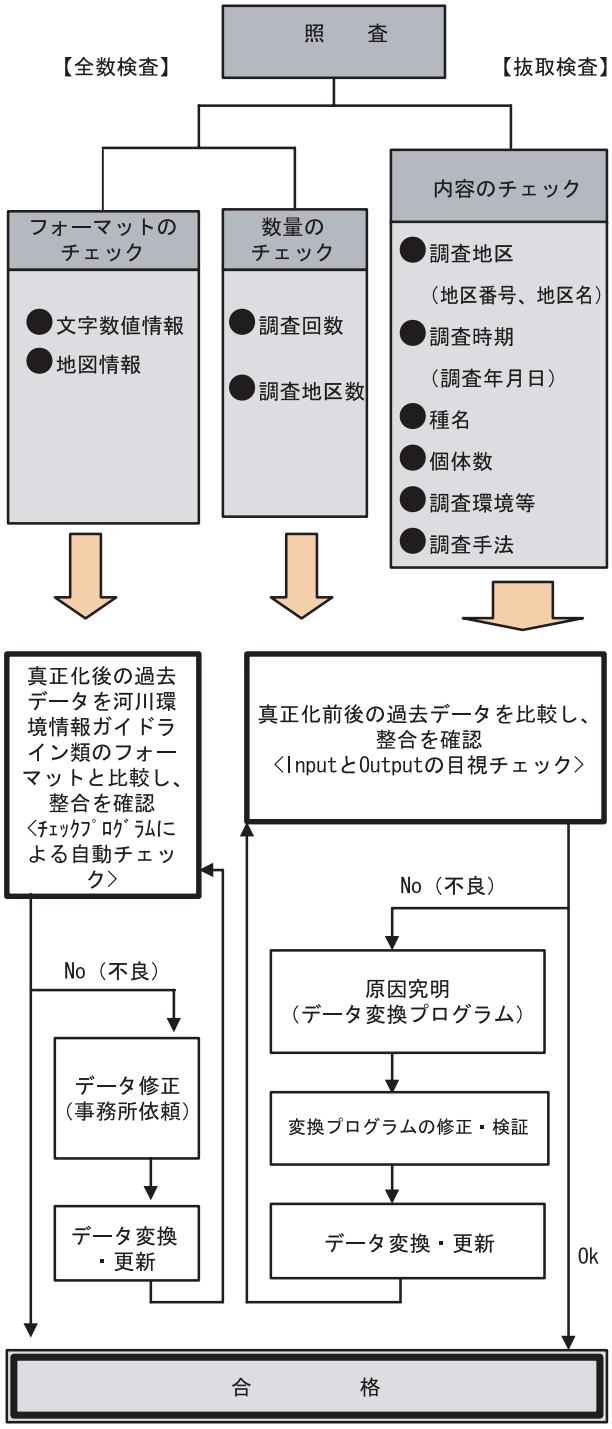
各照査項目に対する検査の方法は、表－４に示すとおりである。フォーマットチェックでは「文字数値情報」及び「地図情報」についてチェックプログラムを開発し、これにより自動チェックを行った。また、数量チェックでは「調査回数」及び「調査地区数」について、内容のチェックでは「調査地区」、「調査時期」、「種名」、「個体数」、「調査環境等」、「調査手法」について目視によるチェックを行った。

表－４　照査の内容

照査項目	検査項目	検査の方法
数量のチェック	調査回数・地区数	調査回数・地区数について整合性を目視確認する
フォーマットのチェック	文字数値情報 地図情報	文字数値・地図情報について、チェックプログラムを用いて所定のフォーマットに合致するか確認する
内容のチェック	調査地区	地区番号及び地区名について整合性を目視確認する
	調査時期	調査年月日について整合性を目視確認する
	種名	確認種のリストについて種名変更履歴を加味し整合性を目視確認する
	個体数	個体数について種名変更履歴を加味し整合性を目視確認する
	調査環境等	流速・水深等の調査条件について整合性を目視確認する
	調査手法	調査手法について整合性を目視確認する

（7）照査

データの照査は、前述の通り全てのデータについて照査を行う全数検査と、全データから検査対象を抜取り抽出する抜取検査の２通りの照査を行った。図－８に作業フローを示す。



図－8　照査のフロー

照査の結果、データ変換プログラムに関するエラーは見られなかったが、手作業を加えた点については若干の修正ミスを発見した。このように不良が確認され

た場合は、その原因を究明した上で対応策を検討し、データの修正後に再チェックを行い真正化前後でデータに不整合がないことを確認した。

なお、抜き取り検査で確認された不良については、その原因が抜き取り検査対象外のデータについても影響を及ぼしていると考えられる場合は、それらデータについて修正を行った。

3．まとめ

以上の作業を終え､過去データの真正化が完了した。当初は、種名の真正化に最も時間がかかると考えられたが、実際に作業を始めると、全データをデータベース形式に変換する作業に最も多くの時間を費やすことになった。このように一度に大量のデータをデータベース化する場合には、なるべく異質なデータを出さないようにすることが重要である。河川水辺の国勢調査のような全国規模のデータは、データ入力者が多数に及ぶため、データの形式やファイルフォーマットに関してエラーが生じやすい。そのようなエラーを防ぐためには、データ形式等を規定する詳細なルールの設定とそれをフォローする体制の整備が不可欠である。

今回は過去データ入力マニュアルがルールに相当するが、数回の調査マニュアルの改定に対応するため、フォーマットや記述内容が複雑になったことが、エラー多発の一因と考えられる。またそのようなエラーに対応するため「ヘルプデスク」を設置し入力者の質問等を受け付けたが、入力者が独自に判断してデータを作成する場合があった。ヘルプデスクは有効な手段の一つであるが、入力者からの積極的なアクセスが無い場合は問題の把握が困難である。このような事態に対しては、例えばチェックシステムを開発・公開し、入力者自らがデータチェックができるようにするなど、データについて入力段階から一定の質を確保することが重要である。

因みに現在の河川水辺の国勢調査の入出力システムでは、入力の段階毎にチェックがかかるため、入力者はエラーを解決できなければ先に進めないようになっている。また、入力データの最終的なチェックシステムが搭載されており、それに合格したデータのみ収集するため、データ収集の時点では「データベースに格納ができない」等の大きな問題はほとんど発生しない。これらの機能は、多くのデータを統一的に取り扱いたい場合（データベース化等）には有効である。

一方で、一連の作業により平成５年度からの河川水辺の国勢調査データが一つのデータベースに格納され、データの検索、経年的な分析が可能になった。現在、河川水辺の国勢調査の結果は、インターネット上で国土交通省の「水情報国土データ管理センター」から「河川環境情報データベース」として平成12年度の結果のみ公開されている（注：平成15年６月現在)。今後、今回真正化した過去データが同様に公開されることにより、これまで年鑑で４割程度だった情報量がほとんど全て公開され、一般市民あるいは研究者に河川環境の基礎情報として様々な場面で広く利用・活用されるのではないかと考える。

〈参考文献〉

1）林尚・小川鶴蔵：河川環境情報の効率的な整備と有効活用について「リバーフロント研究所報告第13号」財団法人リバーフロント整備センター(2002)