被爆者データベースにおける個人同定処理

共通機器部門情報基盤機器管理班
原　憲行

1. はじめに

原爆放射線医科学研究所附属被ばく資料調査解析部の所有する被爆者データベース（Atomic Bomb Survivors Database：ABS データベース）は，広島で原爆に被爆した人のうち，広島県及び広島市の被爆者健康手帳を取得した人物を対象集団とした包括的なデータベースである．2019 年 1 月 29 日時点で登録されているデータは295,036 名 3,280,781 件となっている．

データベース構築にあたっては、広島市被爆者票，原爆被災復元調査等の複数の資料を基にしている．各資料から作成されたデータには，個人を特定する唯一の番号（ABS 番号）で紐付けされている。個人の二重登録を避けるためには、各資料のデータ同士を紐付けする処理（個人同定処理）が欠かせない。

今回は主に，新規データ追加時等に行う個人同定処理について報告する．

2. 個人同定処理

データベースに新しいデータを追加する場合，それがまったく新しい個人のデータなのか，それとも登録済みの個人への追加データなのかが重要となる．個人同定処理は，一致するデータ項目の数によってこの区別を行う．処理は，コンピュータによって自動的に行われる第一段階と，人手による第二段階に分けられる．

(1) 第一段階

「姓名（漢字，カナ）」「住所」「生年月日」「性別」による比較を個人単位で行い，「同一人と見なすもの」「別人とするもの」「保留」の 3 種に分類する．比較結果の例を表 1 に示す．

	姓名漢字		姓名カナ		住所	生年月日				性別
	姓	名	姓	名		西暦年	月	日	和暦元号	
同一人物	○	○	○	○	○	○	○	○	○	○
	○	○	○	○		○	○	○	○	○
保留			○	○	○	○	○		○	
	○	○	○	○	○	○			○	○
	○	○				○	○		○	○
別人						○	○	○		○
		○		○	○					
						○	○			

表 1 個人同定処理第一段階　比較例

値が一致した項目に「○」をつけている．値が一致しない場合，および，データがその値を持たず比較そのものが不可能な場合は空白になっている．各項目の値はいくつかの要因から複数の値を取り得るために，一致した項目数だけでは比較結果を唯一に分類できない．あらかじめ重視する項目を指定しておくことが必要になる．経験上，性別と元号は他人同士でも一致しやすいため，「姓名」「住所」「生年月日」が優先されやすい．

(2) 第二段階

前段階で「保留」と分類した群に対して，「被爆町」「入市町」「家族情報」等による比較を行う．データ化されていない部分を比較するために原医研所有の文書資料を参照することもある．

(3) 誤判定に備えて

個人同定処理には 2 つの誤りが生じ得る．一つは，同一人物を別人と判定する誤り．もう一つは，別人を同一人物と判定する誤りである．どちらもデータベースとしての精度を落とすものだが，ABS においては後者を特に避けるよう意識しており，疑わしいものは別人として判定している．これは，別人同士とみなしたままであれば，その後に追加されたデータが決め手となって同一人物になり得るが，一度同一人物と見なした場合は，その後別人と判定し直すきっかけ

が生じにくいからである.

(4) 判定結果

基となる資料が十分な情報を持っている場合，個人同定処理は第一段階だけで完了できる. しかし実際は，姓名が漢字もしくはカナの一方だけしかないため，あるいは代理記入による姓名や生年月日の揺れ等のために，本当に同一人物であってもすべての項目が一致するとは限らない. 例として，2016 年の死亡者に関する個人同定処理の内訳を表 2 に示す. 姓名カナについては含まれていないため，それ以外の項目を比較した結果，4031 組の候補が出来，その内 17 組が別人であった.

姓名漢字	住所	生年月日	元号	性別	候補数	同一人物
○	○	○	○	○	1282	1282
				○	1	1
○		○	○	○	2743	2726
			○		1	1
				○	4	4

表 2　2016 年死亡情報　判定結果

最後に，別人とみなされていたものを同一人物と見直した例を示す. 本来ならば同一人物となるはずの 2 つのデータの一方のみが死亡データと結びついた場合，他方は延々と「生存」扱いの不自然なデータとなる. そこで，そのようなデータの候補として「大正元年以前に生まれ」「手帳交付記録がなく」「死亡データもない」こと等を条件に ABS から抽出した 3234 名を対象に，同定処理の第二段階を行った結果，27 組の新たな結合が見つかった. 発覚した組における項目の一致度合いを表 3 に示す. 姓または名の不一致は表記揺れの範疇といえるものが多かったが，生年月日の不一致は年月ともに大きくずれているものがあった.

姓名漢字	姓名カナ	住所	生年月日	元号	性別	新結合数
○	○			○	○	2
○	○				○	1
○				○	○	1
	○	○		○		1
	○			○	○	4
		○	○	○	○	1
		○		○	○	3
			○	○	○	1
				○	○	12
				○		1

表 3　死亡なし新結合　一致度合い

3. 結び

ABS においては，データの基となる資料の年代や目的が異なるため，機械的に同一人物と判定するための基準を定めるのは困難である. さらに、人手による作業は，多大な時間を要するという欠点がある. このため、いかにして精度を損なうことなく判定の自動化を行うかが，今後の課題の一つと言える.