データフュージョン技術を活用したデータリッチ化への取り組み

片柳　伊佐

ビデオリサーチでは，データフュージョン技術によって，パネル調査データ同士の融合や調査データによる大規模データの拡張を実現し，顧客サービスに活用している．調査データに特有な課題や，大規模データとのフュージョンを実現するための対応策と検証結果，および実例を交えて，これまでのデータフュージョンへの取り組みについて紹介する．

キーワード：データフュージョン，調査データ，大規模データ

1.　はじめに

ビデオリサーチは，テレビの広告取引指標となる視聴率を測定する会社として 1962 年に設立された．テレビ番組や新聞で視聴率が引用される際は「ビデオリサーチ調べ」という言葉が必ず付くため，目にされたこともあるかもしれない．

設立から半世紀を経て，現在では視聴率調査以外にもさまざまなメディア調査を行っている．テレビに関しては番組評価調査や CM のクリエイティブ評価調査，タレントイメージ調査などを継続的に実施しており，テレビ以外についてもラジオや新聞，雑誌，Web の接触状況を把握する個別調査を続けている．

また，こうしたメディア調査に加えて，生活者を横断的に捉えるシングルソース調査である ACR/ex（エーシーアール エクス） [1] や，テレビ視聴と Web 接触を共に機械式で測定している VR CUBIC などのマーケティング調査も行うなど，多種多様な調査データを取り揃えている．

こうした調査をシングルソース化し，一人の対象者になるべく多くの項目について調査・測定できれば生活者について広く深く捉えることができるが，それには四つの点で限界がある．まず，あまりに調査項目が多いと「対象者の負担」が重くなり，回答慣れや飽きを起こし正確な回答を得られなくなる可能性がある．次に「調査コスト」が挙げられる．負担の大きな調査への協力を対象者から得るには，費用と時間の両方のコストがかかることになる．そして一度に多くの項目を調査しようとすると，一つの調査項目が別の項目の回答や行動に影響を与える「調査バイアス」を招くリスクがある．最後に，継続的に行っている調査の場合は特に，経年変化を捉えるために「時系列性の担保」が重要であり，調査結果に影響を与えかねない項目の追加や入れ替えは容易にはできないという実態がある．

このように，調査におけるシングルソース化に限界がある中で，ビデオリサーチでは単独の調査ではカバーしきれない項目をデータフュージョンによって付加し，調査データの活用場面を広げることに取り組んできた．

2.　汎用的データフュージョンシステムの開発

データフュージョンとは，独立して取得した二つのデータを，サンプルの共通項目の類似度に基づいて融合し，疑似的にシングルソースデータを作り出すことである．

ビデオリサーチでのデータフュージョンの取り組みは，2003 年からスタートした．既に実施済みの二つの調査データをフュージョンするという試みで，フュージョン実行に必要なパラメーターなどもそれぞれの調査データに特化したものだった．つまり，特定時点の特定調査データ同士に最適化されたフュージョンということである．

これに続いて 2007 年からは，NTT データ数理システムの協力も得て，調査時点・調査内容に依存しないデータフュージョン手法の開発に取り組み始めた．このときに採用したのは「制約付き統計的マッチング」と呼ばれる手法である．一方，制約を付けないマッチング方法もあり，以後本稿では「制約なし統計的マッチング」と呼ぶ．

二つの手法の違いに入る前に，データフュージョンの概念説明において使われる「レシピエント」と「ド

かたやなぎ　いさ
株式会社ビデオリサーチ　ソリューション事業局ソリューション開発部
〒 102–0075 東京都千代田区三番町 6–17

＜レシピエントデータ＞

標本番号	性別	番組視聴
a1	1	1
a2	1	0
a3	2	1
		66.7%

＜ドナーデータ＞

標本番号	性別	商品所有
b1	1	1
b2	2	0
		50.0%

＜制約なし統計的マッチングによるフューズドデータ＞

標本番号	性別	標本番号	番組視聴	標本番号	商品所有
f1	1	a1	1	b1	1
f2	1	a2	0	b1	1
f3	2	a3	1	b2	0
N=3			66.7%		66.7%

図 1　制約なし統計的マッチング

ナー」という言葉の定義をしておく．これらの単語は一般的に臓器移植に関して用いられ，その場合，ドナーは臓器を提供する人を指し，その臓器を移植される人がレシピエントと呼ばれる．この提供する人，される人という関係がデータフュージョンにおいてもそのままあてはまり，二つのデータを融合する際に項目を提供する側のデータがドナーで，その項目が融合されるデータがレシピエントとなる．

「制約なし統計的マッチング」においては，レシピエント一人に対して，類似度が最も高いドナーが一人マッチングされるのが基本形である．ドナーから見て類似度が同じレシピエントが複数いる場合は，一人のドナーが複数のレシピエントと紐づくことになる．

非常に単純化した例を図 1 に表した．ドナーが 2 サンプル，レシピエントが 3 サンプルの調査データである．ドナーデータでは「性別」と「ある商品の所有有無」を取得しており，商品所有率は 50.0%となっている．レシピエントデータでは「性別」と「あるテレビ番組の視聴有無」を取得しており，視聴割合は 66.7%である．

この二つのデータを性別の類似度でマッチングすると，性別が 1 のドナー b1 が，同じ性別の値をもつレシピエントの a1 と a2 の 2 サンプルにマッチングされる．同様に，性別が 2 のドナー b2 は，レシピエント a3 と紐づけられる．こうしてできあがったフューズドデータでは，サンプル b1 が 2 サンプル分に膨らむため，商品所有率を集計すると 66.7%になる．つまり，「制約なし統計的マッチング」の結果，データフュージョンをする前のドナーデータから商品所有率が変化することになる．

一方の「制約付き統計的マッチング」は，ドナーデータでの商品所有率 50.0%という値を保持するようにサンプル同士をマッチングする手法である [2]．そのために，ドナー，レシピエントそれぞれがサンプル間の類似度に基づくウエイト値を付与され，多対多でマッチングされる．

図 1 と同様のデータを例に，「制約付き統計的マッチング」を表したのが図 2 である．ドナーデータとレシピエントデータの集計値を保持するという制約の下にマッチングを行うため，「制約付き」と言われる．図 2 において性別の類似度でサンプル同士をマッチングすると，ドナー b1 はレシピエント a1, a2 の 2 サンプルとマッチングされるが，a1 と b1 の組み合わせはウエイト 2 を，a2 と b1 の組み合わせはウエイト 1 を付与される．ドナー b2 についても同様で，レシピエント a2, a3 の 2 サンプルとマッチングされ，a2 と b2 の組み合わせがウエイト 1 を，a2 と b3 の組み合わせがウエイト 2 をもつことになる．

マッチング後にフューズドデータを集計する際にウエイトを加味することで，商品所有率はドナーデータと同じ 50.0%となる．

ビデオリサーチが行っている調査の多くはランダムサンプリングで実施されており，そのデータには市場の代表性がある．そうしたデータを用いてデータフュージョンをする際に，集計値がデータフュージョン前後で変わることは望ましくない．そのため「制約付き統計的マッチング」を採用した．

2003 年の実施時は特定の二つの調査データに特化したデータフュージョンだったのに対して，2007 年からの取り組みでは，フュージョン対象データのサンプル間の類似度を算出する式を，データに合わせて都度生成できる仕組みを目指した．つまり，どのようなデータを投入しても対応可能な，汎用的なデータフュージョンシステムを開発したのである．そのベースエンジン

＜レシピエントデータ＞

標本番号	ウエイト	性別	番組視聴
a1	2	1	1
a2	2	1	0
a3	2	2	1
			66.7%

＜ドナーデータ＞

標本番号	ウエイト	性別	商品所有
b1	3	1	1
b2	3	2	0
			50.0%

＜制約付き統計的マッチングによるフューズドデータ＞

標本番号	ウエイト	性別	標本番号	番組視聴	標本番号	商品所有
f1	2	1	a1	1	b1	1
f2	1	1	a2	0	b1	1
f3	1	1	a2	0	b2	0
f4	2	2	a3	1	b2	0
	N=6			66.7%		50.0%

図 2　制約付き統計的マッチング

としては，NTT データ数理システムの汎用数理計画法パッケージ Numerical Optimizer を使っている [3].

開発したシステムの特徴を整理すると，以下三つになる.

- 制約付き統計的マッチングを採用
- サンプル間類似度の算出式を都度生成
- レシピエントとドナーのペアごとにウエイト値をもつ多対多マッチング

こうした機能をもつ汎用的なシステムを開発したことで，社内や顧客からデータフュージョンの要請があった際に適宜対応できるようになった.

3.　サンプルローテーションへの対応

2007 年に開発した汎用的フュージョンシステムを使う中で，サンプルローテーションへの対応という調査データ特有の課題が新たに見えてきた.

サンプルローテーションとは，同じ調査対象者に定期的に調査を行うパネル型調査において，サンプルの一部を少しずつ入れ替えることを指す．たとえば，調査開始時に 1,000 人の調査パネルを敷き，3 か月ごとにそのうちの 100 人を調査対象外として，新たに 100 人を調査対象者に加えることを繰り返す．調査対象外になったサンプルをローテーションアウトサンプル，新たに調査に加わったサンプルをローテーションインサンプルという．サンプルローテーションの目的には，サンプルを定期的に入れ替えて代表性を保ちつつ，一度に多数のサンプルが入れ替わることで集計データの時系列性に影響が出ることをなるべく抑えるという側面もある.

前述のとおり，汎用的データフュージョンシステムでは，フュージョン後も調査データの元の値を維持するために「制約付き統計的マッチング」を採用した．しかし，サンプルローテーションがある調査データの場合，「制約付き統計的マッチング」だと課題がある.

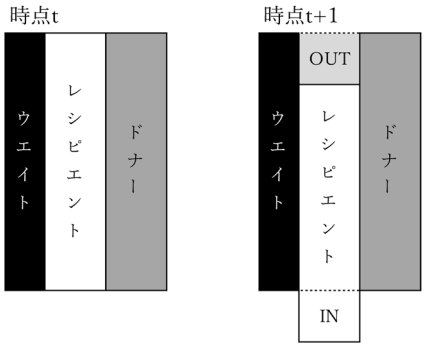


図 3　制約付き統計的マッチングによるフューズドデータとサンプルローテーションの関係

図 3 にあるように，時点 t で「制約付き統計的マッチング」でデータフュージョンを行うと，ウエイトを加味することで元の集計値を維持できるフューズドデータになる．ところがその後，レシピエントデータでサンプルローテーションが行われると，時点 t+1 には一部のレシピエントがローテーションアウトしていなくなっている．それらのレシピエントとマッチングされていたドナーは，マッチング相手がいなくなってしまう．また，レシピエントデータで新たにローテーションインしたサンプルには，マッチング相手となるドナーサンプルがいない状態になる．つまり，時点 t で作成されたサンプルマッチングが崩れてしまうため，「制約つき統計的マッチング」の本来の目的である元のデータを維持するということができなくなる.

よって，サンプルローテーションのある調査データで「制約付き統計的マッチング」によるデータフュー

ジョンを運用していくには，元の集計値を維持するために，ローテーションがある度に全サンプルをマッチングし直す必要がある．これは，サンプルローテーションが頻繁に起こる調査にとっては，運用負荷の増加につながる．また，全サンプルを毎回マッチングし直すためにマッチング相手がその都度変わり，時系列変化を追いにくくなることも課題となった．こうした課題に対応するために，2015 年から 2016 年にかけて新たなフュージョン手法を検討し，従来の汎用的データフュージョンシステムに機能を追加搭載することにした．

最も大きな変更としては，従来の「制約付き統計的マッチング」に加えて，「制約なし統計的マッチング」も可能にしたことである．サンプル間類似度を算出した後に，「制約付き統計的マッチング」では類似度に基づいて多対多マッチングとウエイト値算出を行うが，追加した「制約なし統計的マッチング」の機能によりレシピエントと最も類似度が高いドナーとの 1 対 1 のマッチングができるようになった．

また，既に作成済みのサンプル間類似度の算出式を使って，サンプル間類似度を計算できるようにもした．これにより，サンプルローテーション後の時点 t+1 において，時点 t で作成したサンプル間類似度算出式を使って，ローテーションインサンプルについてのみフュージョンを実施できるようになった．時点 t で生成された式を活用するというのは，基点となる時点 t のレシピエントとドナーの関係性を，ローテーションインサンプルにも適用することになる．

前述のとおり，「制約なし統計的マッチング」では，ドナーデータ，レシピエントデータの集計値は完全には維持されない．しかし，「制約付き統計的マッチング」と同じプロセスで生成された類似度算出式を使うことで，集計値の再現性が高くなることも検証で確認した．

拡張した機能の特徴 3 点は，以下のとおりである．

- 制約なし統計的マッチング
- 既存のサンプル間類似度の算出式でサンプルマッチングが可能
- レシピエントとの類似度が最も高いドナーとマッチングする 1 対 1 マッチング

4. データフュージョンによる視聴率データの拡張

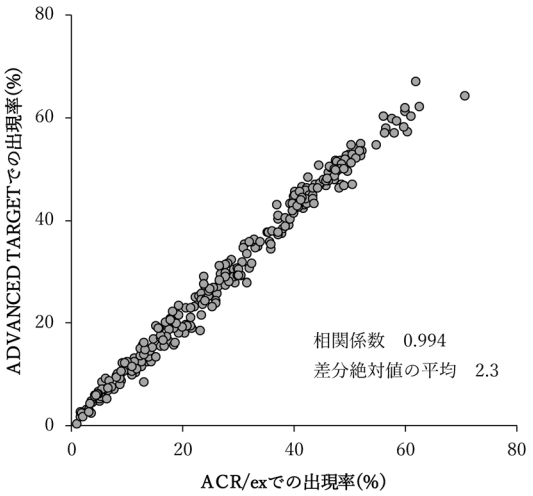
拡張した機能を使って実用化したのが「ADVANCED TARGET」というサービスである．生活者の属性や商品関与，メディア接触などを網羅的に調査したマーケティング調査 ACR/ex をドナーとし，テレビ視聴率調査をレシピエントとしてデータフュージョンすることで，従来よりも細かい生活者属性や嗜好での視聴率分析を可能にした．

これらの二つの調査により得られたデータは，ランダムサンプリングによる代表性のあるデータである．テレビ視聴率調査では，個人のテレビ視聴状況を 1 年 365 日，1 分単位という非常に細かい粒度で測定しているが，対象者個人のプロフィールは性別，年齢，未既婚，職業など主要なデモグラフィックにとどまっている．一方で ACR/ex は，テレビ接触は 1 年の中で特定 1 週間（関東地区では年 2 回 1 週間ずつ）という限られた期間の測定，測定単位も 5 分単位とやや粗くなる．しかし，個人プロフィールについては年収，学歴をはじめ，生活意識や情報感度といった非常に細かい生活者属性から，商品カテゴリ・ブランド別の利用状況や，テレビ以外にもラジオ，新聞，雑誌，Web との接触や交通機関の利用実態など，約 1 万項目ともいう膨大な量の項目を捕捉したシングルソースデータになっている．

テレビの視聴率調査対象者は，ほぼ毎日，全国どこかでサンプルローテーションが起こっているほど頻繁にサンプルの入れ替えがある．そうしたデータに対して，サンプルローテーションに対応して機能拡張したデータフュージョンを行うことで，視聴率調査サンプル全体に ACR/ex の多様なプロフィール項目が紐付けられ，これまでにないテレビ視聴率集計が行えるようになった．

ADVANCED TARGET では，実運用を見据えて「制約なしマッチング」を採用しているが，その場合も元の集計値をほぼ維持できることを確認している．図 4 は，ACR/ex の調査項目約 300 個について，ドナーデータである ACR/ex での出現率と，「制約なし統計的マッチング」でのデータフュージョンによってできた ADVANCED TARGET での出現率を比較したものである．

300 項目は商品カテゴリ関与や年収，テレビ嗜好など幅広い項目にわたっており，点の一つ一つが，たとえば「世帯年収 800 万円以上」「ビール購入決定者」「ドラマ好き」といった項目を示している．「制約なし統計的マッチング」であっても，ドナーの ACR/ex とフューズドデータの ADVANCED TARGET での出現率は，相関係数 0.994，差分絶対値の平均が 2.3 と非常に近く，ドナーデータでの値をほぼ維持できていることがわかる．



サンプル数：ACR/ex=4,877, ADVANCED TARGET=1,465

図 4　制約なし統計的マッチング前後での調査項目約 300 個の出現率比較

ADVANCED TARGET で可能になった視聴率集計の一例として図 5 を示す．左の図は，「個人全体（4 才以上）」がどの時間帯によくテレビを視聴しているかを表しており，ここまでは従来の視聴率データでも集計可能だった．それに対して，右の「自家用乗用車　購入決定者」は，視聴率に ACR/ex をデータフュージョンし，視聴率調査サンプルに商品カテゴリ購入関与度が推定付与されたことでできた集計ターゲットである．色が濃い時間帯は視聴率がより高いことを示しており，「個人全体」と「自家用乗用車　購入決定者」では視聴状況に違いがあることがわかる．

このように，新たに開発したデータフュージョン手法が，ADVANCED TARGET というサービスの実用化につながり，またそのサービスにより今までできなかった切り口でのテレビ視聴率の集計や分析が可能になった．

5.　大規模データのデータフュージョンへの対応

調査データ同士のデータフュージョンに関しては，課題に対応しながら機能を拡張し，ADVANCED TARGET に見られるように実用化にまで至ったが，次に大規模データのデータフュージョンという新たな課題が出てきた．

インターネットの広がりや，生活のあらゆる面でのデジタル化が進んだことで，さまざまな記録がデータとして蓄積されるようになっている．そうした大規模データの多くは，調査のようにあらかじめ利用目的を定め，設計して取得しているものではない．そのため，単独のデータでは得られる項目が少なく，活用に限界がある場合が多い．そうした課題への対応策としては，「ID 共通化」と「推定プロフィール付与」の二つが挙げられる．

ID 共通化とは各社で個別にデータを集めるのではなく，同じ ID で管理してデータを連携させるというやり方で，さまざまな業種で横断して使える共通ポイントサービスがその典型例である．そのほかにも，異なる Web サービス間で登録・利用時のログイン ID を共通化することや，Web サイトアクセス時に発行される cookie を複数サイト間で統合する cookie-sync も含まれる．

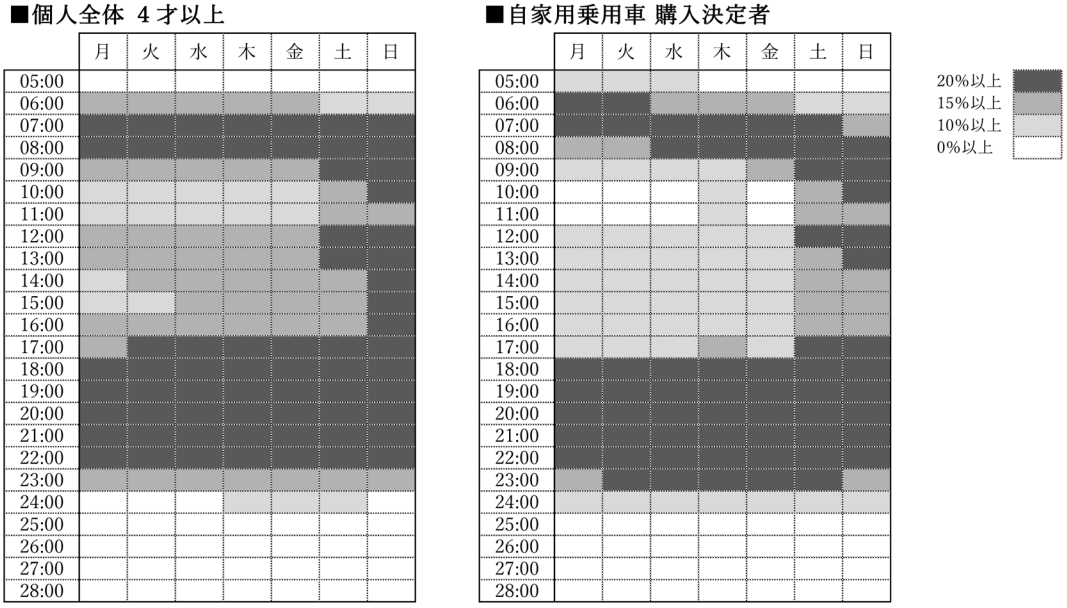
一方で個別に集めたデータに対しては，取得できていない項目を推定で付与するという方法がある．ビデオリサーチには，さまざまな調査データがデモグラフィック情報をきちんと取得したうえで蓄積されているため，調査データを活用して大規模データへ属性を推定付与することに取り組んでいる．

推定でプロフィールを付与する方法はさらに「推定モデルの構築・適用」と「データフュージョンによるデータ融合」という二つに大別される．

推定モデルを活用するには，調査データの中でプロフィールを推定するモデルを作成し，大規模データに推定モデルを適用する．たとえば，Web アクセス履歴とサンプル属性をシングルソースで取得している調査データを使用し，Web アクセス履歴から性別・年齢・職業といったデモグラフィックや，所有商品や興味関心カテゴリを推定するモデルを構築する．そのモデルを顧客が所有する数千万ユーザーのアクセス履歴に適用し，推定属性を付与するサービスを展開している．推定付与された属性データは，主に広告配信に活用されている．

データフュージョンを使ってのデータ融合では，調査データをドナー，大規模データをレシピエントとし，サンプルマッチングにより調査データの属性項目を大規模データに丸ごと紐付ける方法を取る．こちらは，前述の ADVANCED TARGET で実用化した手法を活用している．ただ，従来のシステムでは大規模データのデータフュージョンに対応できない部分があったため，2017 年にフュージョンシステムのさらなる機能拡張を行った．

まずデータ数に関して，これまではレシピエントデータ，ドナーデータの件数に 1 万件ずつという上限があった．調査データ同士であればほぼこの限度内に収まっ



データ：ビデオリサーチ テレビ視聴率調査（関東地区、個人視聴率）
率区分：全局毎60分平均視聴率
期間　：2020/6/29(月)～2020/8/2(日)

図 5　ADVANCED TARGET により可能になった視聴率集計事例

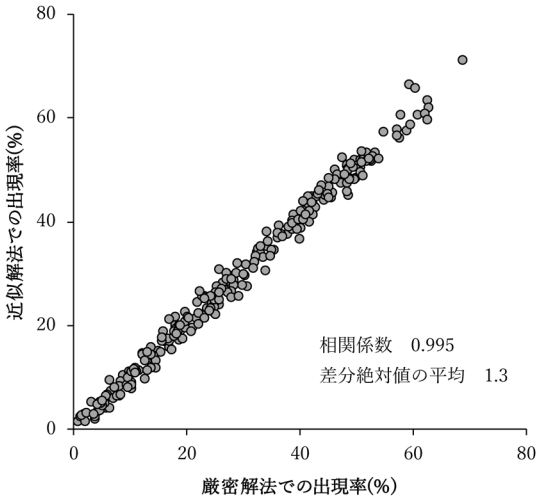
ていたが，大規模データは 1 万件を軽く超えることが多いため，処理可能なデータ件数を数百万件規模まで拡大した．

また，データ規模が大きくなると，フュージョンプロセスに時間がかかるという問題がある．その対策として，サンプル間類似度の算出式はデータをサンプリングして作成するという機能を追加した．

加えて，サンプルマッチングと，「制約付き統計的マッチング」のウエイト算出において，従来は厳密解法で行っていた部分を，大規模データフュージョンではスピードを優先して近似解法で行うこととした．近似解法に変えると，制約式は満たすが最適なマッチングとなる保証がないため，結果にどう影響するかも検証している．

図 6 は，同じデータを使って，「制約なし統計的マッチング」によるデータフュージョンを厳密解法と近似解法でそれぞれ実施し，前述の ADVANCED TARGET での確認に用いたのと同じドナー項目 300 個について，フューズドデータでの出現率を比較した結果である．相関係数 0.995，差分絶対値の平均が 1.3 となり，厳密解法と近似解法でのフュージョン結果が非常に近い値になっていることが確認できる．

大規模データ対応として拡張した機能をまとめると，以下の 3 点である．



サンプル数：厳密解法=1,660, 近似解法=1,660

図 6　厳密解法と近似解法での調査項目約 300 個の出現率比較

- 処理できるドナーとレシピエントのデータ件数が数百万件レベルまで拡大
- データをサンプリングしてサンプルの類似度算出式を作成
- サンプルマッチングとウエイト算出を近似解法で実施

こうして拡張した新機能も，既に実務で活用されている．顧客のもつ大規模な会員データや行動データに対して，調査に基づくプロフィールや意識・行動データを付与するという使い方である．ビデオリサーチが蓄積している各種の調査データを大規模データとフュージョンすることで，大規模データの利用機会や利用シーンの拡大につながっている．

6. おわりに

2007 年の汎用的データフュージョンシステムの開発から始まり，サンプルローテーション対応および大規模データへの対応という実務での課題に応じて，機能を拡張してきた．データフュージョンの対象も，調査データ同士から，調査データと大規模データと広がってきている．そして，機能拡張するにつれて，データフュージョンシステムの利用頻度や利用案件も増えてきている．

社会のあらゆるところでさまざまなデータが蓄積されている中で，データフュージョンはデータの利活用をさらに進める有用な技術である．今後も実務での課題に応じて，さらなるシステムの機能拡張と技術の活用を進めていきたい．

参考文献

[1] ビデオリサーチ，ACR/ex, https://www.videor.co.jp/service/communication/acrex.html（2020 年 9 月 23 日閲覧）

[2] R. Soong and M. de Montigny, “An anatomy of data fusion,” In *Paper for the Worldwide Readership Research Symposium*, Venice (Italy), pp. 87–109, 2001.

[3] NTT データ数理システム，Numerical Optimizer, https://www.msi.co.jp/nuopt/（2020 年 8 月 31 日閲覧）