The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

データ 3.0 時代のデータランドスケープ

Data Landscape in the Era of Data 3.0

早矢仕晃章*1
Teruaki Hayashi

大澤幸生*1
Yukio Ohsawa

*1 東京大学大学院　工学系研究科　システム創成学専攻
Department of Systems Innovation, School of Engineering, The University of Tokyo

The potential expectations for discovering problems and solving them by combining data and knowledge in different domains have been increasing. However, there are many social barriers to its realization. Data is originally the property that expresses the observed phenomena or events in the world in symbols and character strings, which is Data 1.0 in this paper. Afterward, Data 2.0 has been developed by the privilege of personal devices or Internet of Things, which is the era of linked data stored in different domains. While the benefits of data 2.0 are immeasurable, the technologies and discussions to combine only data are reaching the limits. In this paper, we propose "Data 3.0" as a new stage in the cross-discipline data utilization and discuss the Data Landscape in the era of Data 3.0.

1. はじめに

近年，様々な分野で蓄積された膨大なデータを利用し，意思決定に役立てようとする動きが活発になり，ビッグデータの世界的なブームも起きた．さらに，人工知能（AI）の普及も後押しし，データの役割はますます重要となってきている．そのような社会背景の中で，既存の分析技術や統計学では扱えないデータが登場してきた．さらに今まで AI が主要技術ではなかった分野においても AI 関連技術に対する注目が集まっている．そこで，一つの企業や組織のデータではなく，異なる領域のデータや知識，AI 技術を流通・交換・連携させることで新しい価値を発見し，問題解決を行う動きが進展してきた．

本質は変わらないものの，データを取り巻く環境，役割，価値は時代と共に大きく変容してきた．原初のデータは，世の中で起こる複雑な現象を理解するために，対象を観察し，特徴を抽出し，符号化したものであった．すなわち，データの本質は，事実を文字，数値，記号として記述したプロパティであると言える．だが，ビッグデータのように計算機やセンサーの高度化によって，膨大かつ高粒度なデータが取得可能となった．このような技術の進展によって，蓄積されたデータの再利用によって付加価値を生み出したいというニーズが高まり，様々な技術やサービスが誕生した．さらに，異なる領域の異種のデータを繋げることにより，新しい発見を促す技術も発展してきた．その中で，Web 上でデータを交換するプラットフォームを事業化したデータ市場の一形態も登場してきた．そして，AI 技術の社会実装により，異なる領域のデータを組み合わせるだけでなく，知識との融合によって問題を発見し，解決を目指す動きが萌芽し始めた．以上のように，人類の歴史の中で登場した初期のデータと比較し，現在のデータは質や量のみならず，意義，利用価値，役割が大きく異なっていると言えよう．

連絡先：早矢仕晃章，東京大学大学院　工学系研究科　システム創成学専攻，hayashi@sys.t.u-tokyo.ac.jp
大澤幸生，東京大学大学院　工学系研究科　システム創成学専攻，ohsawa@sys.t.u-tokyo.ac.jp
This research was supported by JST, CREST. 本研究の着想を得るにあたり，2018 年 3 月に退官された東京大学大学大学院工学系研究科システム創成学専攻教授　大橋弘忠先生の最終講義を参考にさせて頂きました．ここに感謝致します．

本稿では，異分野のデータ連携と知識の融合の時代におけるデータを「データ 3.0」と定義し，異分野データ連携の課題とデータ 3.0 時代におけるデータの在り方とデータがもたらす世界の見え方（データランドスケープ）について論ずる．

2. データの役割とその変遷

2.1 データ 1.0 からデータ 2.0 へ

原初のデータとは，文字，数値，記号として記述可能な事実であり，分からない現象を理解するために用いられてきた．例えば，太陽や星の位置を記録することで時間や年月を把握したり，様々な物理量に単位系を与えることで共通理解を得ることなどが挙げられる．つまり，データは実世界の複雑な現象を解釈可能とするために，分析対象の特徴を切り出し，仮想世界の中で対象を再構成するために誕生した．データを用いることによって，人間は実世界及び社会の事象を処理し，理解し，知識を伝達することができるようになった．これがデータの始まりであると筆者らは考える．つまり，この黎明期におけるデータは，伝達，解釈，処理に適する形式（文字，数値，記号）に表したものであり，「データ 1.0」と呼べる段階であったと言えよう．データ 1.0 におけるデータの扱い方は分析（analysis）であり，すなわち「ある事柄の内容や性質を明らかにするため，細かな要素に分けていくこと」であった．

続く「データ 2.0」はデータの統合（synthesis）の時代となる．データ 1.0 におけるデータ分析は，物事を理解するために，より細かく局所的に事象を観察し，データ化すること物事の内容や性質を明らかにすることが主であった．しかし，分野が細分化し，取得可能なデータも多岐に渡り，変数同士の組み合わせも膨大となる中，「分析」という語は従来の語義である「細かな要素に分けていく」作業だけでなく，異なる領域のデータを結合することで物事の内容や性質を理解する試みも含むようになった．つまり，対象とする事象だけでなく，対象事象の周辺で起こっている事象を含めて，大局的に見ることで大きな物事の関連性を発見することが求められるようになった．データ 2.0 と呼べるこの潮流は，データ融合（data fusion）や，オープンデータ，Linked Data [Berners-Lee 06], Linked Open Data などのセマンティック Web 技術や社会的な運動の原動力となった．オープンデータ

の盛り上がりによって，行政が保有するデータに誰もがアクセス可能となる社会に対する期待も高まった．データ 2.0 はデータの公開と民主化の時代でもあったと言えよう．

2.2 データ 3.0 におけるデータ市場

データ 2.0 の時代には，Web の発展，スマートフォンなどのパーソナルデバイスの普及やセンサーの高度化，Internet of Things（IoT）の進展などによって様々な領域のデータを連結させた結果，分野横断的なデータ駆動型イノベーションが起こり，多くの成功を収めた．しかし，データ 1.0 の時代と比較してさらに多様なデータが取得され，結合されるようになった結果，データモデルが複雑化・多様化した．もちろん，全ての人がデータについて詳しいわけではない．誰もがアクセス可能なデータのネットワーク（知識ベース）を作っても，データの構造を把握している人，あるいは設計者しか欲しいデータを検索・発見できなくなってしまった．しかも，ビッグデータのブームに代表されるように，データの設計者でさえ全体像を把握することが困難なほどにデータは膨大となった．データ 2.0 時代のデータは人間の認知の限界を超えてしまった．例えば，関連分野の論文でさえも全てを読破することがほぼ不可能となったことが挙げられる．さらに，有益な分析結果を得るためには，単一のデータソースだけでなく，複数のデータを適切に組合せることが重要[Ellram 16]であるが，データの組み合わせは無数にあり，全ての組み合わせを考慮することはもはや不可能である．行政や自治体では，オープンデータ戦略に則り，データを公開したもののなかなか利用されていないという問題も存在する[狩野 18]．データ 2.0 によって，データが広く社会のシステムに普及した一方で，システムの複雑化によって逆にデータ保有者とデータ利用者の溝が深まってしまったとも言える．

データ 2.0 がもたらした恩恵は計り知れないものの，上述のように，データのみを結合する技術及び議論は限界を迎えつつある．そこで，本稿ではデータ利活用における新たなステージとして，「データ 3.0」を提案する．データ 3.0 の社会では，データだけでなく，データと人間の知識が結びつき，問題の発見と解決が行われることが期待される．データ 3.0 時代におけるデータは，従来のデータ 1.0 及びデータ 2.0 におけるデータとは，性質が異なる．データ 1.0 ではデータは未知の現象を理解するためのものであった．そしてデータ 2.0 では，データ同士を繋げることで新しい価値や知識が生み出されてきた．続くデータ 3.0 のデータは，人間の知識と結合することで人と人を繋ぐ役割を果たすものである[1]．すなわち，データ 3.0 によってデータを介して人と人が出会い，既存の領域では扱えない問題を発見し，知識の連携によって異業種共創が行われる社会となると考えられる．

筆者らは，現在の社会はデータ 2.0 からデータ 3.0 への過渡期に位置すると考えている．日本では，2014 年にデータ駆動型イノベーション創出戦略協議会が創設され，民間ではデータエクスチェンジ・コンソーシアム等が設立された．これらの取り組みの目的は単純にデータを結合することではなく，異分野データ連携による人の邂逅である．つまり，データにはプライバシーの問題，データの用途の不透明性，ビジネス機会の損失などの問題が内在しているため，通常は秘匿にされることが多いが，これらの取り組みはデータを介して人と人が出会い，共創する環境の創造が目的であるという点で，データ 2.0 におけるデータとは性質が異なることが分かる．さらに 2016 年には経済産業省と民間企業が協業し，異なるデータ事業者が連携できるデータ流通市場の事業化に着手している．Web をプラットフォームとするデータの市場の誕生もデータ 3.0 に向かう一つの社会的な動きであると見ることができる．すでに，Microsoft Azure，Qlik，KDnuggets，Factual，Infochimps といった，データを売買・交換するサービスが登場してきている．また，データ提供者とデータ利用者のコミュニケーションによるデータの価値化を促す仕組みとして，個人のオンラインショップの購買履歴や健康情報などを一括管理する「情報銀行（Information Bank）」という仕組みも実現しようとしている [秋山 13, 砂原 14]．情報銀行とは，パーソナルデータを取り扱うハブとなる組織であり，そこに各個人が情報を預け，集積された情報を活用し，情報を預託した個人に何らかのメリットを返す仕組みを提供する．「銀行」というメタファーを用いているのは，各利用者が，自分が預託した情報を的確に把握できるようにし，それらの情報の利用方法の把握することを可能とするためである．さらに上述のコンソーシアムや協議会の興隆に加え，日本データ取引所[2]がデータの取引を仲介する市場を開拓し，EverySense 社[3]がセンシングデータの流通マーケットプレイスなど展開し，データ流通推進協議会[4]がデータの運用方法の策定及びプラットフォーム整備を進めているように，異なる分野のデータを取引きするための場を創生することにより，データの流通を促進させる仕組みがデータ市場の一形態として社会実装されてきている．以上に概観したように，異なる領域のデータを組み合わせることで新しい価値を発見し，分野を横断して知識と結合させて問題発見と解決をするデータ 3.0 が萌芽し始めている．

しかし，データ 2.0 からデータ 3.0 に完全に移行できたわけではない．データを交換し，取引きするデータ市場が展開されているが，それらには様々な問題が存在する．Web をプラットフォームとする現在のデータ市場は，データの表層的な情報を Web 上に陳列するだけに留まっており，ステークホルダー間のコミュニケーションによるデータの価値化と，イノベーションの場としての市場の機能が有効に働く環境としてはまだ十分ではない．なぜなら，データの表層的な情報を列挙しただけでは，データ提供者とデータ利用者の間での利用方法の提案，評価というコミュニケーションの活性化による知識の交換に至るのは難しいからである．また，利用者とのコミュニケーションが欠如していれば，データ保有者も自身が保有するデータの価値を理解する機会を得ることができない．データの価値は利用文脈に大きく依存すること[Hayashi 17b]を考慮すると，データの適切な価値付けが行われないだけでなく，利用価値が不明確なデータは，市場というプラットフォームに登場することができない．そのため，データを連結し入手可能な状態にするだけでなく，人間の知識と結びつけることで価値あるデータを選び，入手する交渉や熟考を伴う検討の場としてのデータ市場が必要なのである[大澤 17]．すなわち，変数によるデータの結合だけでなく，共通の目的や知識によってデータを繋ぐ必要がある．データ 3.0 時代のデータ利活用と異分野データ連携を促すためには，変数の結合だけでなく，データを結びつける人間の知識やデータの利用文脈をも結合することが重要となる．

[1] データ 3.0 と比較し，データ 1.0 及び 2.0 が劣っているというわけではないことに注意したい．データが変容したのではなく，データを取り巻く社会と人が変化したことがデータ 1.0 からデータ 2.0，そしてデータ 3.0 に推移しつつあることの本質である．

[2] http://j-dex.co.jp/

[3] https://every-sense.com/

[4] https://data-trading.org/

3. データランドスケープ

データは主として変数と値によって構成されている．データ 2.0 において，データの連結はデータの変数を揃えることによって達成される．つまり，共通する変数を有しているデータ同士は結合可能性が高いと言える．しかし，データ 3.0 においては，データと人間の知識の融合によって異分野の共創が実現することはすでに述べた．本章では，それらの違いをデータランドスケープとして俯瞰的に分析する．

図 1 は 1098 件のデータジャケット（Data Jacket: DJ）[Ohsawa 13, Hayashi 17a]を用いて異なる領域のデータを「変数」と「文脈」によって連結したネットワーク（データランドスケープ）である．異分野データ連携によるデータ駆動型イノベーションを実現するためには，世の中に存在するデータの構造とそれらの関係を正しく理解する必要がある．つまり，個々のデータを分析するのではなく，データによって構成されるデータの母集団がどのような構造的特徴を有しているのかを調べることが重要である．データランドスケープは，データが作り出す世界の傾向や特徴を定量的に評価するためにデータの関係を俯瞰し，知識を体系付けるための可視化である．

データジャケット（Data Jacket: DJ）は，データ自体を秘匿としたままデータの概要情報を記述するためのフレームワークである．データの概要情報とは，データに関する説明文，含まれる変数の名前，保存形式，共有条件などを意味する．つまり，DJ はデータのデータ，すなわちメタデータの一形態である．しかし，通常のメタデータは機械可読性を高めるための記述方法であるが，DJ は人間がデータについて理解し，議論可能とすることを目的としている．この手法により，データの中身は公開できなくとも，どこにどのようなデータが存在し，どのような情報を有しているのかおよそ把握可能となる（図 2）．また，異なるフォーマット，異なる粒度の変数ラベルを有するデータを共通の記述ルールによって構造的にメタデータ化することで，様々な形式のデータを統一的に扱うことができる．変数ラベルとは，データ固有の変数を自然言語によって記述された説明文を意味する．

データランドスケープでは，各ノードは DJ を表し，DJ 同士が共通する変数ラベルを保有している場合は青色のリンク，共通する文脈を持っている場合は赤色のリンクが張られるものとする．例えば，「東京都の病院の位置情報」と「東京都に設置されている警察署の設置情報」という DJ が「緯度」と「経度」という変数ラベルを共通して有していた場合，2 つの DJ には青色のリンクが張られることになる．また，同データが「東京都における位置データ」という文脈を共有している場合，2 つの DJ は赤色のリンクで結ばれる．なお，文脈は DJ に含まれるデータの概要説明（データ概要）の説明文から単語を抽出して用いた．

図 1 は DJ を用いてデータランドスケープの可視化を行ったものである．図 1 の上図は変数の繋がりから表現されたデータランドスケープと文脈の繋がりから表現されたデータランドスケープの合成ネットワークである．表 1 に DJ を用いたデータランドスケープのネットワークの特徴量を比較した．「変数」は変数によるネットワークを表し，「文脈」は文脈によるネットワーク，そして「変数+文脈」は変数及び文脈によるネットワークを表す．表にあるように，文脈によるネットワークの平均次数は 33.46 と高く，リンク密度 0.030 とやや高い．一方，変数のみの繋がりによるネットワークでは，平均次数は 20.57，リンク密度は 0.019 であった．つまり，文脈によるネットワークの方が密なネットワークを作りやすい性質を持っていることが分かる．一方，変数によるネットワークは同類選択性が 0.586 と高く，コンポーネント数が 374 個と多いことから，同じようなデータは互いに密なネットワークを構成するものの他のデータとは繋がりは疎となる．一方，文脈によるネットワー

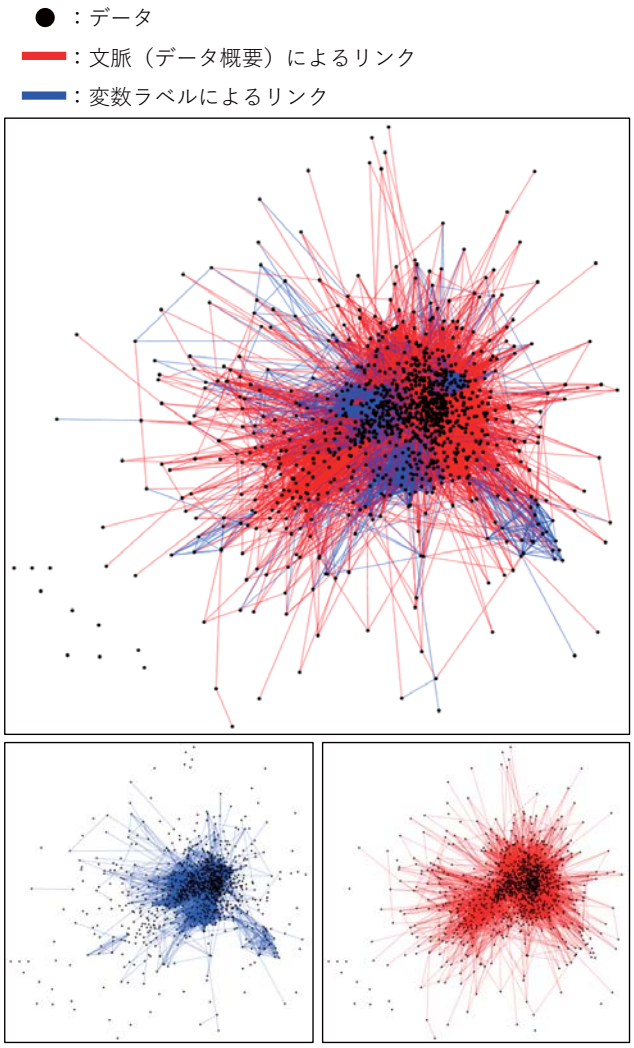

図 1　DJ を用いたデータランドスケープ（上図：変数と文脈による合成ネットワーク．下図左：変数によるネットワーク．下図右：文脈によるネットワーク）

DJ No. XX【購買履歴データ】	
概要	東京都の◯◯スーパーマーケットで収集されている顧客の購買行動履歴。
収集方法・コスト	ポイントカードとPOSによって取得
共有条件	共有不可
データの種類	表形式、テキスト、数値
保存形式	CSV
分析・シミュレーション	時系列分析
変数ラベル	氏名、性別、顧客ID、支払金額、購入品目、日にち
分析結果	• その日の売上の計算 • 今後の売上の予測と仕入れの推定
期待される分析	顧客の購買行動とリピート率を計算し、ロイヤルカスタマーの特定が可能かもしれない。
コメント	有効なデータの組み合わせが発見されればデータの提供あるいはコラボレーションもあり得る。

図 2　DJ の記入例（[早矢仕 18a]より引用）

クは同類選択性が 0.185 と比較的低く，さらにコンポーネント数は 26 個であった．つまり，文脈によるネットワークは局所的に塊を作るのではなく，大局的に密なネットワークを構成することが分かる．さらに，リンク数を見てみると，変数と文脈の合成ネットワークは 28902 であり，変数のネットワークは 11292，そして文脈のネットワークは 18367 である．変数によるリンクと文脈によるリンクの両方で繋がっているデータはごくわずかであり，リンク数は 757 と極めて少ないことが分かる．つまり，変数と文脈の両方が共通しているデータは非常に少ない．そして，この結果によって，変数を介したネットワークだけでは繋がりえないが，共通の利用文脈を補完することによってデータが結合可能となることも示唆される．すなわち，変数によるネットワークがデータ 2.0 のデータランドスケープ（図 1 の青色リンクのみの繋がり）を表し，人間の知識によってデータの共通の利用文脈が発見され融合するデータ 3.0 のデータランドスケープ（図 1 の赤色と青色リンクによって構成されるネットワーク）とすると，データ 3.0 では異分野のデータが密に融合していると言うことができる．

表 1　データランドスケープのネットワークの特徴量

特徴	変数	文脈	変数 +文脈
ノード数	1098	1098	1098
リンク数	11292	18367	28902
平均次数	20.57	33.46	52.64
リンク密度	0.019	0.030	0.048
クラスタ係数	0.465	0.531	0.509
同類選択性	0.586	0.185	0.146
コンポーネント数	374	26	11

4. おわりに

異なる分野のデータと知識を結合することで問題を発見し，問題解決を行うことへの期待が高まっているものの，その実現には多くの障壁が存在する．本稿では，データ 3.0 時代におけるデータの在り方についてデータランドスケープを用いて所見を述べた．“Data is the new oil（データは新しい石油である）”と言われて久しい．確かにデータは石油と同様に意思決定において重要な材であり，ビジネスにおいては経済財であるかもしれないが，石油はそれ自体ただの物質であり，データは記号の羅列に過ぎない．石油は適切な加工を施すことによって価値が生まれる．そして，データは適切な利用文脈が与えられ，分析などに利用されることで，情報となり，人間の意思決定に役立つ知識となる．つまり，データの利用価値を策定するプロセスとは，データに文脈を与える行為に相当する．異なる分野のデータに文脈を与えるプロセスの実現と異業種共創のプラットフォーム[岩佐 18, 佐々木 18, 早矢仕 18b]が生まれてきていることから，データ 3.0 は徐々に社会に浸透しつつあると考えられる．現在はデータ 2.0 からデータ 3.0 への過渡期である．これからのデータ駆動型社会の実現には，データ・AI 技術・人間の相互作用による異分野のデータと知識の連携によるイノベーションの場であるデータ市場の整備が重要と言える．

参考文献

[秋山 13] 秋山寛子, 山内正人, 柴崎亮介, 砂原秀樹: 情報銀行システムにおける個人情報蓄積機構の機能設計と実装，マルチメディア，分散協調とモバイルシンポジウム 2013 論文集, pp.1953-1957 (2013)

[岩佐 18] 岩佐太路，早矢仕晃章，大澤幸生: Web 版 Innovators Marketplace on Data Jackets を用いたデータ利活用法に関するコミュニケーション支援，信学技報，人工知能と知識処理研究会, Vol.117, No.440, pp.55-60 (2018)

[Ellram 16] Ellram, M.L., and Tate, L.W.: The Use of Secondary Data in Purchasing and Supply Management (P/SM) Research, Journal of Purchasing and Supply Management, Vol.22, No.4, pp.250-254 (2016)

[Ohsawa 13] Ohsawa, Y., Kido, H., Hayashi, T., and Liu, C.: Data Jackets for Synthesizing Values in the Market of Data, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science, Vol.22, pp.709-716 (2013)

[大澤 17] 大澤幸生, 早矢仕晃章, 秋元正博, 久代紀之, 中村潤: データ市場, 近代科学社（2017）

[狩野 18] 狩野英司, 大西浩史: 官民データ活用による行政課題解決サイクルの仕組み化に向けて~「官民データ活用シナリオ創 発プラットフォーム事業」の取組み~, 信学技報，人工知能と 知識処理研究会，Vol.117, No.440, pp.67-72 (2018)

[佐々木 18] 佐々木泰芳, 池田栄次:「ブロックチェーン×データジャケット」によるデータ流通・利活用社会への加速，信学技報， 人工知能と知識処理研究会，Vol.117, No.440, pp.43-47 (2018)

[Berners-Lee 06] Berners-Lee, T.: Linked Data – Design Issues, 2006. https://www.w3.org/DesignIssues/LinkedData.html, [最終アクセス 2018 年 3 月 9 日]

[早矢仕 18a] 早矢仕晃章, 大澤幸生: データ市場におけるデータのネットワークと関係性の分析―データの属性と繋がりからの考察―, 信学技報，人工知能と知識処理研究会, Vol.117, No.440, pp.49-54 (2018)

[早矢仕 18b] 早矢仕晃章，大澤幸生：データジャケットを用いた異分野データ連携，人工知能学会誌，AI とデータ-データに基づく意思決定と社会イノベーション創出-特集，Vol.33, No.2, pp.140-148 (2018)

[Hayashi 17a] Hayashi, T., and Ohsawa, Y.: Matrix-based Method for Inferring Variable Labels Using Outlines of Data in Data Jackets, The Pacific-Asia Conference on Knowledge Discovery and Data Mining（2017）

[Hayashi 17b] Hayashi, T., and Ohsawa, Y.: Preliminary Case Study on Value Determination of Datasets and Cross-disciplinary Data Collaboration Using Data Jackets, 21st International Conference on Knowledge Based and Intelligent Information and Engineering System, Vol.112, pp.2175–2184 (2017)

[砂原 14] 砂原秀樹, 山内正人, 金杉洋, 柴崎亮介:「情報銀行」構想とその技術的課題，マルチメディア，分散協調とモバイルシンポジウム 2014 論文集, pp.1024-1026 (2014)