博士論文（要約）

データ市場における知識構造化に基づく
データ利活用シナリオ検討プロセスの研究

早矢仕　晃章

本論文の要旨

本研究の長期的な目的は，データ市場におけるデータ駆動型イノベーション（Data Driven Innovation : DDI）に貢献することである．データ市場とは，データの公開・共有を強制するのではなく，自由市場の原理で利用者が必要なデータを選び，所有者と交渉の末に入手できるプラットフォームである．市場とは提供者と利用者が接し，その相互作用の中で商材の価値を評価し，その評価に合う条件を設定して取引を行うイノベーションの場である．つまり，データ市場はデータを商材として扱い，ステークホルダー間のコミュニケーションによって価値を決定し，取引を行う場である．

近年，異なる領域のデータを組み合わせて新たな知識を獲得し，意思決定に役立てることへの期待が高まっているものの，データの共有及び公開には高い社会的障壁がある．企業の機会損失やプライバシー侵害のリスク，さらにデータの取扱いについての各国の法制度の違いなどの問題がある．このような状況において，データの公開ではなく，市場における交換という戦略によってDDIを推進しようとする様々な形態のデータ市場がWebを中心に萌芽し始めた．しかし，Webのみをプラットフォームとするサービスは，データの表層的な情報を陳列するだけに留まっており，ステークホルダー間の十分なコミュニケーションが期待できず，イノベーションの場としての市場の機能が有効に働いているとは言い難い．また，データの蓄積方法は積極的に議論されてきたが，既存の知識やモデルでは扱えないデータを含むデータ利活用知識の蓄積方法は十分に議論されてこなかった．つまり，データ市場におけるDDIを促進するためには，データ市場の仕組みを理解するとともに，利用価値のあるデータの利用方法を知り，どのような仮説が検証可能かということを議論するプラットフォームと，データ利活用知識を蓄積し再利用するための技術が必要である．

そこで本研究では，DDI創出環境としてのデータ市場の仕組みとデータ利活用方法の検討による価値化プロセスを理解し，提案手法の有用性を評価するため，実際のデータ市場に参画するステークホルダー（実業家，研究者，分析者など）を参加者とした実験的データ市場を設計した．そして，データ市場に関わる様々なステークホルダーの意図や目的が反映される要求とデータ利活用案をデータ利活用知識として構造化し，ステークホルダー間の創造的コミュニケーションによるデータ価値化とデータ市場創出を支援する技術を開発した．

本研究における実験的データ市場の設計には，基礎技術としてデータジャケット（DJ），データ利活用方法検討ワークショップInnovators Marketplace on Data Jackets（IMDJ），データ利活用シナリオ創出手法アクション・プランニング（AP）を用いた．本研究のコア技術であるDJとは，データの中身ではなく，データの概要情報（データ内の変数名，保存形式，収集方法など）を共有し，価値を検討可能にする方法である．個人を識別す

る情報を含む共有不可能なデータでも, DJにすることでリスクを低減させて情報が共有可能となる．例えば，商品の購買履歴データには氏名，性別，支払金額などの個人を識別する情報が含まれるため，一般公開することはできない．しかし，購買履歴データを「氏名」，「性別」，「支払金額」といった変数名としてメタデータ化すれば，個人を識別する情報を秘匿のまま，データに関する情報を共有できる．実験的データ市場におけるDJの利用とは，データ交換・売買以前のデータに対する期待の高さを表し，データの利用価値を保有者に提示することに相当する．つまり，データ利活用方法の提案によってデータの利用価値が定まり, 需要が生じる. その需要によって供給の必要性が生じ，取引のための条件（価格）が調整されるという市場の原理が働く．すなわち，データの概要情報の共有によって今まで秘匿であったデータの価値の評価が可能となる．

まず，本研究では，データ利活用知識の蓄積方法として，DJだけでなく，過去のIMDJにおいて議論された要求・データ利活用案・データの関係をデータ利活用知識として構造化し，検索システムData Jacket Store（DJストア）を実装した．データ利活用知識構造化により, ユーザーが自分と異なる視点を持つ過去のユーザーが考案したデータの使途を発見したり，別の人が考案したデータ利活用案に注目することで意思決定に役立つDJを探し出すことが可能となることを評価実験により示した．以上により，過去に検討されたデータ利活用知識の構造化と再利用は新たな知識獲得に有用であることが分かった．特に，実験的データ市場において既存の知識や情報だけでは解決できない問題に直面した際に，データに関する情報の陳列だけではなく，データ利活用知識の構造化による検索システムが利用価値の高いデータの発見と問題解決を促し, データに対する新たな需要を喚起する可能性が示唆された．さらに実験では，オープン化できないデータほど, 提案者及び利用者にとって問題解決及び新ビジネス創出において有用性が認められる可能性が高くなることが示唆された．また，検索結果においても秘匿データの方がユーザーの興味・関心の度合いも高いことが分かった．以上の結果から，データ市場はオープンデータに代表される公開可能データのみに閉じられた場ではなく, 公開が難しい個人や企業のデータ及びその保有者を巻き込むイノベーションの場として機能し得ることが分かった．

しかし，データ市場はデータの組合せのみを議論する場ではない．続いて，データ利活用に関わる諸要素が意思決定者のデータ利活用方法検討のプロセスにどのように現れるのかという，シナリオ生成プロセスとシナリオの構造化について議論した．事業計画立案時の筆記行動に着目した実験により, 実験的データ市場においてデータに文脈を付与するシナリオ生成プロセスには仮説推論における非単調性が現れることが分かった．また，データの組合せだけでなく，ステークホルダーやリソースといったデータ利活用に関わる諸要素の関連性を考慮した検討が重要であるという示唆が得られた．

以上に得られた知見を元に，データ利活用知識を拡張し，シナリオ創出手法APによって生成されたシナリオの構造化と再利用の仕組みを提案した．そして，ステークホル

ダー表出と関係推定システムResource Finderを実装し，過去に検討されたデータ利活用シナリオの構造化が，文脈によって異なるステークホルダーのシナリオへの関係を推定するのに有効に作用することを実験的に評価した．さらに，新たにデータを取得する意思決定者を支援する変数ラベル推定方法VARIABLE QUESTを提案した．実験では，データ概要の類似性と変数ラベルの共起性を考慮することで，変数ラベルが未知のデータ概要からそのデータに含まれる可能性の高い変数ラベルが推定可能であることを示した．膨大なデータから必要な知識を発見することが困難であるように，データ市場において複数の領域にまたがって存在するデータ，ステークホルダー，変数など，データ利活用に関わる全ての要素を考慮することは難しい．それ故，意思決定者の異なる価値観や多様な背景知識，意図に対応して構造化された知識ベースとそれを検索するシステムが必要となる．データ利活用知識だけでなく，ステークホルダー及び変数ラベルを含むシナリオの再利用により，潜在的なビジネスパートナーや取得すべき変数についての情報をユーザーに提示できる可能性が示唆された．

最後に，実験的データ市場の枠組みを拡張し，データの入手，分析，課題発見とフィードバックという実社会とのインタラクションを含んだ実装的データ市場を観察する二つの応用実験を行った．一つ目の実験では，データ利活用案の創出，分析シナリオの生成，そして実データ分析によって結果を得る一連の過程を観察した．実際のデータ分析は様々な試行錯誤によって結果を得るプロセスであり，実装的データ市場における分析計画と実データ分析の間にはギャップが存在することが明らかとなった．しかし，上流設計部にあたる分析シナリオを十分に検討していれば，比較的評価の高い分析結果が得られる可能性があることが分かった．二つ目の実験では，分析結果を元にした行動によってステークホルダーから新たな情報と有益なフィードバックが得られることが分かった．また，一般に共有できないデータでも，当該データの活用方法を示したシナリオをデータ保有者に提示すれば，範囲を限定してデータの共有が可能となるという重要な示唆を得た．すなわち，行動しながら新しい知識を取り込むことでシナリオを修正し，分析結果を精緻化するプロセスを経ることが，データ市場における新たなステークホルダーの存在を掘り起こし，データの新しい利用方法の発見を促すものと考えられる．

本研究の提案手法により，データ市場において既存の知識やモデルでは扱えないデータの利活用法及び潜在的なステークホルダー，変数ラベルの発見が支援されるとともに，データ利活用知識ベースが更新されることが分かった．また，それらの手法を利用して新たに知識を獲得する意思決定者の行動が改善されることが示された．本研究の提案手法及び得られたデータ市場のモデルによって，データに関する情報及び知識の蓄積が可能となったことで，従来のIMDJ及びAPをデータ市場創出支援技術として大きく改良したということができる．また，本研究の提案手法によって，事業者の新しいデータの発見や異なる事業者とのインタラクションによって新規事業創出が促進されたことが報告されており，データ市場の創出と発展に貢献したものと考えることができる．

謝辞

本研究論文を執筆するにあたり，東京大学大学院工学系研究科システム創成学専攻大澤研究室の大澤幸生教授に心より感謝申し上げます．研究について様々な場面でご指導いただいたことを感謝致します．

私が初めて大澤教授にお会いしたのは，大学 2 年生の冬学期，工学部システム創成学科 SIM コース（現 SDM コース）の動機付けプロジェクトの講義でした．当時初めて，工学的手法により，データに基づいて人間の意思決定を支援するアイデア発想の手法について知り，シナリオという概念について学びました．ビジネスとして社会的にインパクトのあるアイデアについての研究は世の中で多くなされていることを知っていましたが，アイデアそのものではなく，アイデア創出プロセスや人間の認知，データの分析を研究対象とするアプローチが，当時の私にとって非常に新鮮で魅力的であったのを覚えています．その後 4 年生冬学期にて，大澤教授の研究室に配属になり，IMDJ の前身であるイノベーションゲームから人間の創造プロセスについて研究してきました．学士修了後，2012 年 4 月には大学院進学にあたり，ご縁があり大澤研究室にて修士課程の研究をスタートすることとなりました．学士の研究をさらに深め，2 回の国際学会での発表，様々なワークショップへの参加を経て，自身の研究課題であるプロセス研究を進めてきました．2014 年 3 月に修士を修了し，4 月には大澤研究室の博士課程学生としてさらに工学の道を究めようと，データ市場関連研究を開始しました．

博士課程における先生のご指導は修士課程の時とは比べものになりませんでした．そして，大澤先生の自身の考えを形にし，社会や人間を変えるという強い意志と行動力に，私は大きな影響を受けました．それ故，自身が知のフロンティアに立ち，人類を前進させようと絶えず思考し行動を続ける存在が，私の目指す研究者像となりました．また，大澤研究室での研究活動を通し，多くの会議やミーティング，ワークショップなどに参加する機会を頂き，社会・研究の最前線で活躍されている様々な方々にお会いし，議論させて頂きました．大澤先生は研究における私の父のような存在です．ここに，改めて心から感謝致します．

また，ご多忙の中，本研究について丁寧なご査読と審査の日程調整をしてくださり，様々なご意見，ご教示を賜りました山田誠二 国立情報学研究所教授，阿部明典 千葉大学教授，古田一雄 東京大学教授，和泉潔 東京大学教授，ならびに小林肇 東京大学准教授には深く御礼申し上げます．特に，山田先生は私が入会している人工知能学会会長であり，このような形で審査をして頂けたのは光栄の至りに存じます．また，大澤先生が主催され，私がテクニカルファシリテータを務めさせていただいた AI 技術創成ロードマップワークショッ

プでも基調講演をしていただくなど，審査だけでなく自身の研究者としての在り方についてもアドバイス頂きました．また，阿部先生にはご査読以外でも国際学会，ワークショップのご支援や研究を進めていく上での心構えなどをご指導頂きました．そして，古田先生には，修士課程の副査を担当していただいたこともあり，ご縁がありました．また，私の所属していた東京大学のリーディング大学院プログラム GSDM における活動を見守ってくださり，研究生活に対する貴重なご助言，励ましの言葉を頂きました．この場を借りて御礼申し上げます．和泉先生には，3 年前の修士研究の発表と博士課程における自身の研究計画について非常に厳しいご指摘を頂きました．あの時頂いたコメントに研究への愛情と熱意を感じ，自身を研究に駆り立てました．ありがとうございました．また，小林先生からは本研究に対する建設的で有益なご意見を頂戴しました．ここに改めて御礼申し上げます．

博士課程から在籍している東京大学のリーディング大学院プログラム GSDM に様々な支援を頂いたことが，自身の研究の大きな推進力となりました．特に，GSDM の副指導教員の渡部俊也 東京大学教授，松浦正浩 明治大学教授，鎗目雅 東京大学特任准教授には，研究に関する貴重な助言を頂きました．ここに改めて感謝致します．また，城山英明 東京大学教授，鈴木寛 東京大学教授，光石衛 東京大学教授，横野泰之 東京大学教授，丸山茂夫 東京大学教授，西沢利郎 東京大学特任教授，吉川恒志 東京大学特任教授には，自身の研究課題であるデータ市場創出支援技術について GSDM 合宿等で有益なご意見を頂戴しました．そして，華井和代先生，Roberto Orsi 先生，太田響子先生には GSDM の様々な活動，プロジェクト，ワークショップにご参加いただき，研究発表の場や貴重な助言を頂きました．心より感謝致します．そして，Student Inisiative Project のメンバーを始めとした GSDM の仲間には，毎週のように研究やプロジェクトで集まり，切磋琢磨しながら，自身の研究のブラッシュアップをすることができました．ありがとうございました．

GSDM のプログラムでは，国際プロジェクト演習として，およそ 2 ヶ月半，カナダ，エドモントンのアルバータ州立大学（University of Alberta）に共同研究で留学させていただく貴重な機会を頂きました．University of Alberta にてご指導頂きました Randy Goebel 教授に感謝致します．自身の研究を進める中で，Goebel 先生の洞察力，知識，あらゆる物事に繋がりを見出す力に圧倒されたのを覚えています．滞在中，大変熱心に指導して頂いた上，帰国後も気にかけてくださるお気持ちに心から感謝致します．そして，Goebel 先生の研究チームのメンバーである Mi-Young Kim 様（博士），Ying Xu 様，Shazan Jabbar 様には研究について熱く議論しました．また，研究だけでなく，カナダでの生活のアドバイス，サポートをして頂きました．ここに感謝致します．

本研究を遂行するにあたり，株式会社構造計画研究所の皆様には多大なご支援を頂きました．代表取締役社長 服部正太様，秋元正博様，塚本遼太様，高階勇人様，浜井協様，佐藤壮様，北上靖大様，野深裕也様，冨士本大哲様，千種芳幸様，玉田正樹様，上原太陽様，川原眞実加様には，心より感謝を致しております．そして，元経済産業省 商務情報政策局情報経済課の佐々木紀子様，村田正徳様，小柳輝様には，データ駆動型（ドリブン）イノベーション創出戦略協議会からその後のプロジェクトまで，大変お世話になりました．さらに，実社会における実験の実施にあたり，多くの方々から多大なご協力を頂きました．データエクスチェンジ・コンソーシアムのワークショップでは，橋本大也様，上島邦彦様，實川美紀雄様には大変お世話になりました．NPO 法人 ZESDA の皆様には，ワークショップにご参加いただき，貴重のご助言を頂きました．ここに御礼申し上げます．

また，本研究は日本学術振興会特別研究員（DC2）（東京大学大学院工学系研究科）として，JSPS 科研費 JP16J06450 の助成を受けたものです．また，本研究の一部は JST-CREST の支援によるものです．ここに感謝致します．

本研究をこのような形でまとめることができたのは，多くの先生方や研究者の先輩方，実業家の方々の助言と励ましがあったからです．吉村忍 東京大学教授，鳥海不二夫 東京大学准教授には，研究生活に対する貴重なご助言，励ましの言葉を頂きました．新田克己 東京工業大学教授，高間康史 首都大学東京教授，松下光範 関西大学教授，松村真宏 大阪大学准教授には，国際学会，国内学会，そしてワークショップでご一緒させて頂き，研究を進めていく上での心構えなどをご指導頂きました．そして，元東京大学特任教授・NPO 法人医療ガバナンス研究所長 上昌広先生，奈良由美子 放送大学教授，東京家政大学講師 平野真理先生，東京大学医科学研究所 坪倉正治先生，神奈川県立がんセンター 瀧田盛仁先生には，共同研究及びワークショップにて貴重のご助言を頂きました．そして，伊藤孝行 名古屋工業大学教授，小野田崇 青山学院大学教授，産業技術総合研究所 西村拓一先生，矢田勝俊 関西大学教授，山川宏 ドワンゴ人工知能研究所所長，株式会社ディー・エヌ・エー社外取締役・マネックスグループ株式会社社外取締役 堂前宣夫様，（独）科学技術振興機構 嶋田一義様（博士），大和証券 吉野貴晶様（博士），ネットイヤーグループ（株）代表取締役社長 石黒不二代様には，私がテクニカルファシリテータを務めさせて頂いたワークショップにご参加いただき，準備段階から実施に至るまで多大なご支援を頂きました．また，データ流通環境整備検討会では，安念潤司 中央大学法科大学院教授，宍戸常寿 東京大学大学院教授，そして内閣官房の信朝裕行 IT 戦略推進官，堤和弘 参事官補佐には自身の研究課題とアプローチについて社会的・法律的・実業的側面から貴重なご助言頂きました．心より感謝を致します．

国際学会等の出張では会議だけでなく，大澤先生にご紹介頂いた著名な研究者の方々を訪問させて頂きました．パリ第 6 大学の Bernadette Bouchon-Meunier 教授，Jean-Gabriel Ganascia 教授，シンガポールの Kuiyu Zhang 様（博士），King's College London の Peter McBerney 教授, Stanford 大学の Renate Fruchter 先生, Mark Nelson 先生, Inperial College London の Stenphen Muggleton 教授には，ご多忙の中日程の調整をして頂き，私の研究に対して様々なご意見，ご掲示を賜りました．ここに改めて御礼申し上げます．

毎週のゼミでコメントやアドバイスをしてくださった大澤研究室の皆様，ありがとうございました．日頃から大変お世話になりました．特に，大澤研究室の OB である久代紀之 九州工業大学教授，中村潤 芝浦工業大学教授・パナソニック・エクセルテクノロジー（株）取締役・ＡＶＣテクノロジー（株）取締役副社長・ＡＶＣマルチメディアソフト（株）取締役副社長，小俣貴宣様（博士），吉田隆久様，前田雄佐様にはワークショップのアドバイスや研究の論点の整理のポイントなど，非常に多くのことを教えて頂きました．大変感謝しております．また，特に大澤研究室秘書の安田選子さんには事務手続きだけでなく，研究室での生活について様々な指導をいただきました．心より感謝を致しております．

そして，お忙しい中，実験に協力してくださった皆様，本当にありがとうございました．私を支えてくれた家族，友人，そして，ボーイスカウト日本連盟，東京連盟，そして山手地区の皆様に，厚く御礼申し上げます．さらに，私に世界の広さを教えてくれて，研究者を目指す一つのきっかけを与えてくれたボーイスカウトのバングラデシュ派遣プロジェクト，そしてプロジェクトにてご一緒させて頂きました，岡本学様，石橋正彦 麻布大学名誉教授，檀上善夫様，中野充 新潟青陵大学准教授，元 Bangladesh Activity Group 幹事長 古澤孝太様，笹渕賢人様，泊昌史様，星野輝様，Mohammad Atiq Zaman（Ripon）様，Samsul Azad 様，Naz Here 様，Salimul Raman 様，Mamun Khan 様，Diamond Bhuiyan 様，Habib Ullah Hero 様を始め，日本連盟事務局国際部の皆様，バングラデシュ連盟の皆様，韓国連盟の皆様，台湾連盟の皆様，そして，本プロジェクトに関わったすべての方々に感謝致します．

最後に，スカウトとして切磋琢磨するとともに，私の研究生活を支えてくれた婚約者，武田英香里さんに感謝の意を表し，謝辞とさせて頂きます．

目次

第 1 章　序論

1.1 はじめに

近年，データを資源と見ることで，蓄積された膨大なデータを再利用し，様々な分析手法を用いて発見した新しい価値を意思決定に役立てようという動きが活発になってきている．特に，異なる領域のデータを組み合わせて新たな知識を獲得し，意思決定に役立てることへの期待が高まっている．しかし，データの共有及び公開には高い社会的障壁がある．

そこで，データの公開・共有を強制するのではなく，市場の原理で利用者が必要なデータを選び，所有者と交渉の末に入手できるデータ市場が提案され，発展してきた．すでに解決方法が一般的に知られている問題，あるいは一貫性が仮定できる問題であれば，既存のモデルや解決方法を用いれば良いが，データ市場は今まで世の中に出てこなかったデータ，ステークホルダー，知識が登場する新しい市場である．そのため，既存の分析手法を適用することで分析結果が得られるデータもあるかもしれないが，その多くはまだ適用する分析手法が定まらない，または既存の分析手法では扱えないデータである．既存の知識やモデルでは扱えないデータの新しい利活用法及び潜在的なステークホルダーを発見することが求められている．

1.2 研究の目的

本研究の長期的な目的は，データ市場におけるデータ駆動型イノベーション（Data Driven Innovation：DDI）に貢献することである．データ市場とは，データの公開・共有を強制するのではなく，自由市場の原理で利用者が必要なデータを選び，所有者と交渉の末に入手できるプラットフォームである．市場とは提供者と利用者が接し，その相互作用の中で材[1]の価値を評価し，その評価に合う条件を設定して取引を行うイノベーションの場である．データ市場では，データを市場における材として扱い，データの価値と関連知識が話し合われ，外在化され，共有される．すなわち，DDI とは，データ市場というイノベーション創出環境におけるデータの価値化を意味する．

本研究は，データ市場に関わる様々な関係者（ステークホルダー）のニーズとデータ利活用案をデータ利活用知識として精緻化し，ステークホルダー間の創造的コミュニケーションによるデータの価値化を支援する技法の開発を行い，データ市場活性化及び DDI への貢献を目指すものである．本論文では，初めにデータ市場がどのような社会的要請から発展してきたのかを関連研究及び事例を示しながら明らかにする．そして，データ市場にお

[1]「材の金銭的価値が不明な段階を経ながら価値を評価し，その結果として金銭によって評価すべきである場合のみ，初めてその材は「財」という貝偏を持つ漢字を当てるべきものとなる．」（大澤ら，2017）に則り，「材」の字を当てている．

いて解決すべき問題を定義する．続いて，それらの問題に対する本研究の提案手法として，どのようなデータを使うことによりどのような問題が解決可能であるのか，というデータ利活用知識を蓄積し，再利用する仕組みについて議論する．

そして，データ利活用方法検討時の人間の検索行動，筆記行動，議論プロセスに着目し，データ利活用方法を検討する人間の知識獲得とシナリオ生成を支援する検索システムの実装と，その性能評価を行う．本研究によって，既存の知識やモデルでは扱えないデータの利活用法及び潜在的なステークホルダーの発見が支援されるとともに，データ利活用知識ベースが更新される．そして，それらの手法を利用して新たに知識を獲得する意思決定者の行動が改善されていくプロセスが実現することを示す．

本研究の目的を具体的に以下の 5 つにまとめる．

① データ市場の必要性について，既存研究及び事例から論じる

② データ市場創造と活性化の障壁について論じる

③ データ市場創造と活性化に必要な諸技術について論じる

④ 本研究の提案手法とその評価について論じる

⑤ 本研究が明らかにした点を踏まえ，今後の研究と発展可能性について述べる

1.3 本論文の構成

本論文は，7 章から構成される（図 1-1）．

第 1 章では，本研究の背景及び目的について記す．

第 2 章では，データ利活用とデータ市場に関する本研究の背景の詳細及び先行研究について述べる．

第 3 章では，本研究の着眼点とアプローチについて説明し，本研究の提案手法の基本的なアイデアについて述べる．

第 4 章では，本研究のコア技術であるデータジャケットと，それを用いたデータ利活用方法検討ワークショップ Innovators Marketplace on Data Jackets について説明し，データ利活用知識の構造化と検索システムの実装と実験について述べる．

第 5 章では，シナリオ生成手法アクション・プランニングについて説明し，シナリオ構造化による支援システムの実装と実験について述べる．

第 6 章では，第 4 章及び第 5 章を踏まえ，実装的データ市場におけるデータ利活用プロセスについて，提案手法を用いて実社会において行われたデータ利活用の応用実験とその結果について述べる．

第 7 章では，本論文の総括と本研究の成果を結論とし，今後の研究の展開について述べる．

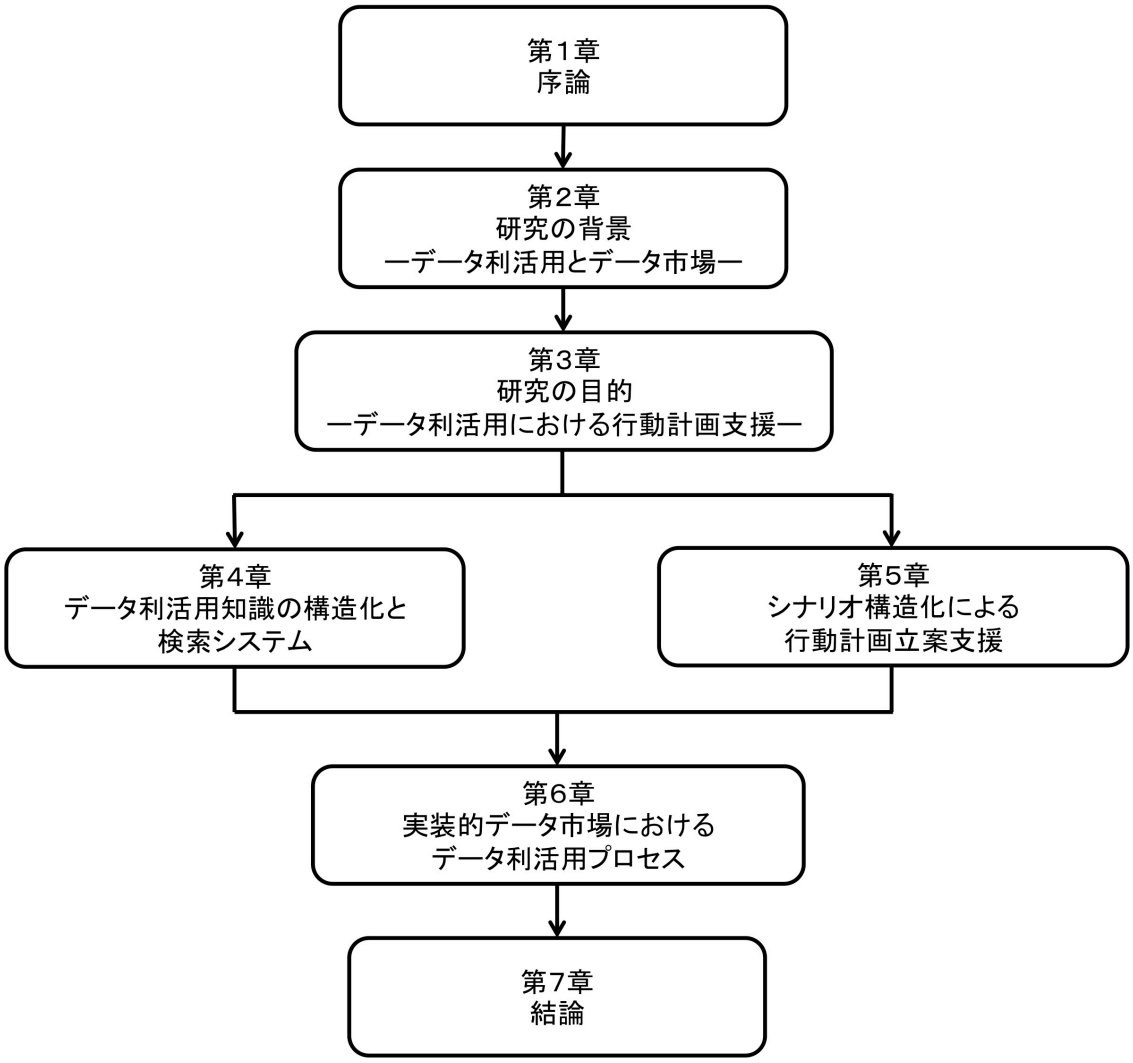
第1章
序論
第2章
研究の背景
ーデータ利活用とデータ市場ー
第3章
研究の目的
ーデータ利活用における行動計画支援ー
第4章
データ利活用知識の構造化と
検索システム
第5章
シナリオ構造化による
行動計画立案支援
第6章
実装的データ市場における
データ利活用プロセス
第7章
結論

図 1-1　本論文の構成

第 2 章　研究の背景 –データ利活用とデータ市場–

本章は 5 つの節で構成されている．まず 2.1 節にてデータ，情報，そして知識の違いについて述べ，データの実世界における役割と位置づけについて認知科学のモデルを用いて議論する．そして，2.2 節において分野を横断したデータ利活用への期待について，具体的な事例を用いて説明する．2.3 節では，2.2 節で説明したデータ利活用における現状の課題と踏まえ，本研究が対象とするデータ市場がどのような社会的要請から発展してきたのかを関連研究及び事例を示しながら明らかにする．そして，2.4 節にて，データ市場の創出支援技術の必要性について議論し，データ市場において解決すべき問題を定義し，それに対する解決アプローチを示す．そして，2.5 節にて本章のまとめを行う．

2.1 データ・情報・知識

有史以来，人間は様々なデータを取得し，利用し，意思決定に役立ててきた．まずは，データ，情報そして知識について本研究における定義について，先行研究を概観しながら明確にする．そして，データの利用価値の策定方法とその市場性について先行研究を概観しつつ説明する．

村上（1980）は，人間の事実の認識（fact）は外界に存在する情報（data）の取捨選択という行為によって創り出されるとした．つまり，人間が外界を理解し，知識を得るという行為には，必要な情報の取捨選択というフィルタが存在している可能性を示唆している．林ら（Hayashi et al., 2006; 林ら, 2007）は村上（1980）の例を用い，同じ物理的実態（外界に存在する情報）を観察する 2 人の問題解決者は，それぞれの持つ知識や置かれたバックグラウンドなどの違いが作り出す視点の差異によって異なるフィルタを作り出すことを指摘した．すなわち，異なるフィルタを通して対象となる data を観察することで，両者は同一の data を観察しているにもかかわらず，異なった事実を構成し得ることを実験的に明らかにした．つまり，人間の外界の事象の認識は，人間の背景知識，事象に対する期待や感情などによって形成されるフィルタに依存するということが理解できる．また，経済学においても Metcalfe（1998）は，経済における意思決定者は個々が最も合理的な選択をして行動しているが，その根拠は背景知識（knowledge）や機会（available opportunities）に依存するため，同じ世界を見ていても，異なる世界を認識していると，認知科学的アプローチと同様の指摘をしている．

一方，Boisot & Canals（2004）は人間の事実の認識及び外界に存在する情報について，経済学から物理学の視点から，刺激（stimuli），データ（data），情報（information），知識（knowledge）としてさらに詳細にモデル化した．データとは，物理世界（world）の刺激をある知覚可能な範囲で取得したものであるとしている．つまり，知覚フィルタ（perceptual filters）を通し

て取得した物理世界の刺激がデータとして取得される．そして，データは概念フィルタ（conceptual filters）を通して情報に変換される．知覚フィルタ及び概念フィルタは，エージェントの認知的・感情的期待（cognitive and affective expectations）によって調整されるものである．この認知的・感情的期待が外界から得られた刺激を認識する人間のフィルタに影響することは Clark（1997）と Damasio（2000）の先行研究でも指摘されている．これらをまとめたものを図 2-1 に示す．また，Boisot & Canals（2004）はデータ，情報，知識をそれぞれが異なる有用性を有した経済財（economic goods）として扱えることを指摘した．経済財とは，稀少性があり，対価を支払うことで消費できる財（商品）を意味する．つまり，データの有用性は物理世界に関する情報を導出できるところであり，情報の有用性は知識の期待や知見を得られるところである．また，知識の有用性はエージェントが適応した方法で物理世界において行動可能にするところである．

さらに，Boisot & Canals（2004）は，データは事実や事象を表す客観的プロパティであるのに対し，情報は文脈や関係に依存するプロパティであると指摘している．Bateson（1972）が「情報とは，差異をもたらす差異の集まり（a difference which makes a difference）」であると述べたように，データから得られる情報は個人や個々の環境，歴史的背景，感情及び気質（Damasio, 2000），そして個々の持つメンタルモデル，そして状況における個人の期待に依存するのである．

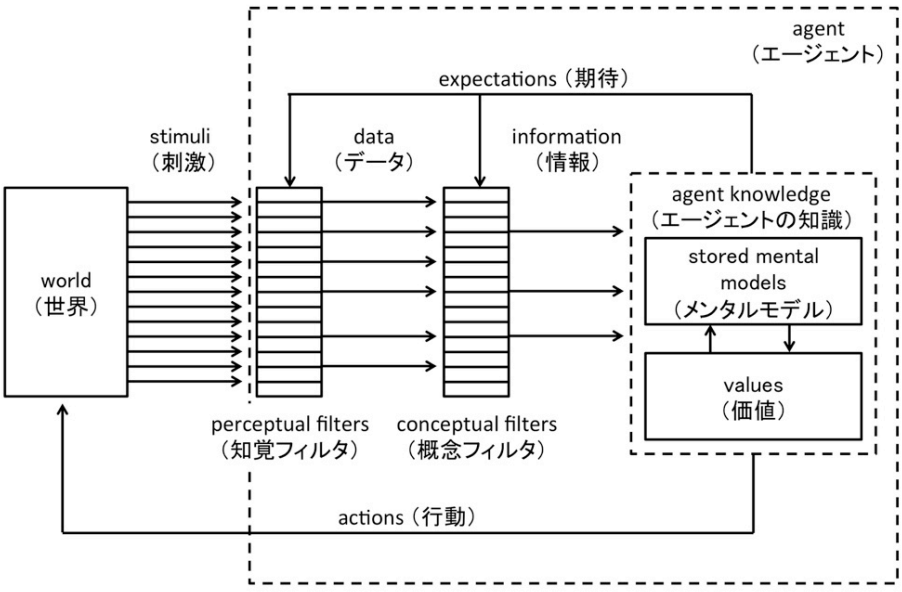

図 2-1　The agent-in-the-world （Boisot & Canals, 2004，筆者一部改変）

Sterman（2000）はデータから得られる情報は文脈に依存することの興味深い例を示している．1974 年にオゾン層の破壊が指摘されたが，多くの科学コミュニティはその事実に懐疑的であった．しかし，1985 年に南極のオゾンホールの存在が証明されたことをきっかけに，NASA の科学者は再度データを精査した．すると，NASA は 1978 年からオゾン層の破壊を示すデータを取得していたのである．つまり，当初からオゾン層の破壊を示すデータを取得してきたにも関わらず，そのような現象が存在することが認識されていなかったため，情報として得られていなかったということがあった．この有名な例は様々な文献で引用されているが，西村（2003）は，この例は「認識できなければ行動できない」という示唆を含んでいると述べ，個人が持つメンタルモデルが人間の世界に対する認識に影響を与えていることを指摘している．すなわち，適切な文脈が与えられなければ，データから情報を抽出し，人間の行動を促す知識とすることが困難であることを示している．

以上を考慮し，本研究におけるデータ，情報，知識の定義は以下である．

(1)　データ：文字，数値，記号として記述可能な事実

(2)　情報：データに文脈が与えられたもの

(3)　知識：データ・情報を利用し，人間を行動可能とするもの

データは文字，数値，記号として記述可能な事実であり，情報はデータに文脈が与えられたものであると定義すると，データとは的確な利用文脈が発見されていないものであると言うことができる．つまり，データ市場においてデータとは未だ価値が見出されていない材を意味すると言える．以上の議論を踏まえると，データから得られる意思決定において重要な情報はエージェントの背景知識や世界を認識するフィルタによって異なるということができる．すなわち，データの価値は利用目的やステークホルダーによって異なるのである．

本節ではデータ・情報・知識を認知モデルから説明し，データの利用価値の策定には文脈の与え方が重要であることを述べた．次節では，異なる領域のデータを活用することで意思決定に役立てようとする実社会における期待とアプローチについて説明し，産業界，学術界におけるデータ駆動型イノベーションの動向及び課題について述べる．

2.2 データ駆動型イノベーション

2.2.1 データ利活用への期待

近年，データを資源と見ることで，蓄積された膨大なデータを再利用し，様々な分析手法を用いた新しい価値の発見や意思決定に役立てようという動きが活発になってきている．特にビッグデータというキーワードが流行し，分野を横断したデータの組合せと利活用によって，既存のサービスの付加価値向上や新製品の開発に対する期待が高まってきている．特に，マーケティング分野でビッグデータが注目されるようになったのは，ブログや facebook，twitter などの SNS（ソーシャル・ネットワーキングサービス）による個人の情報発信が盛んに行われるようになったことが要因の一つであると考えられる．スマートフォンなどのパーソナルな情報端末の普及により，今まで取得困難と言われてきた個人の購買行動履歴などが取得可能となった．従来のマスマーケティングから，「個」に特化したマイクロマーケティングが脚光を浴び，様々なサービスが考案されてきた．

また，ビッグデータが注目される以前から，行政においてはオープンガバメント盛り上がりを見せ始めていた．オープンガバメントとは，行政のデータを一般的にアクセスできる状態にすることで組織の透明性を高め，市民の行政への参加を促す運動である（経済産業省, 2009）．その運動の一環として，二次利用を許可した形で行政のデータを公開し，再利用を促すオープンデータという仕組みが普及してきた．米国の DATA.GOV[2]，英国の DATA.GOV.UK[3], そして，日本の DATA.GO.JP[4]では，このオープンガバメント戦略に則り，行政のデータを様々な形式で公開している．さらに現在，データを Web 上に公開するだけでなく，意味やつながりを構造化し再利用性を向上させるため，オープンデータの RDF（Resource Description Framework）での提供が行われるようになってきている（大向, 2013; 武田, 2011; Berners-Lee et al., 2001; Berners-Lee, 2006 など）．行政のオープンデータを連結することで分野を横断したデータの Web を作り，集合知の生成（Auer et al., 2007; Arndt et al., 2015 など）や推薦システム（Ristoski et al., 2014 など）の開発を行う LOD の研究が行われてきている．全国に先駆けて「データシティ鯖江」としてオープンデータへの取り組みを始めた福井県鯖江市[5]は，RDF によるデータ公開とアプリケーションの提供を行っている．

また，製造の現場では今まで取得されてきた膨大なデータに加え，センサーや計器の高度化によって，物，機械，人間の行動や自然現象が時々刻々と生成している高粒度で精密なデータが取得できるようになった．さらに，これらの情報をウェアラブルデバイスなどのセンサーで取得し，インターネットを介して様々な場所で活用しようとする The Internet

[2] https://www.data.gov/
[3] https://data.gov.uk/
[4] http://www.data.go.jp/
[5] http://data.city.sabae.lg.jp/

of Things（IoT）と呼ばれる動きも注目されている．健康関連事業や労働環境の改善などへの応用が期待され，IoT の潜在的な経済効果が試算されている（Manyika et al., 2015）．以上のように，センサーの高度化とデータ同士の結合により，膨大なデータが日々蓄積され，利活用されてきている．

経済産業省ではデータ駆動型（ドリブン）イノベーション創出戦略協議会（経済産業省, 2014a）が創設され，民間では企業間のデータ交換による新事業創出のための場としてデータエクスチェンジ・コンソーシアム[6]（DXC, 2014a），オープンデータ／ビッグデータ利用推進フォーラム[7]，Open Fog Consortium[8]など様々な業界においてデータ利活用及びマネジメントを目的とした団体が設立された．これらの協議会やコンソーシアムの設立からも分かるように，政府や大学のみならず，産業界においてもこうしたデータに基づく意思決定とデータ利活用への期待が高まりを見ることができる．保有するデータを用いた既存事業の付加価値向上や異なる分野のデータを組み合わせることで新ビジネスを創出したいという潜在的な期待が高まってきていると言えよう．今ではビッグデータという言葉自体は下火となりつつあるものの，データ利活用による成功例は様々なメディアで取り上げられ，多くの組織がデータ利活用に対して意欲的になっている．

2.2.2 データ共有及び利活用の現状と問題

データ利活用とデータに基づく意思決定への期待の高まりがある一方で，データの交換や利活用については，様々な問題が指摘されている．特にプライバシーの問題とデータの用途の不透明性がリスクと考えられている．Acquisti & Gross（2009）はデータの複製の容易さから，個人データの組み合わせは深刻なプライバシー侵害を引き起こすことを指摘した．また，Xu et al.（2014）の研究では，様々なステークホルダー（data provider, data collector, data miner, decision maker）の目的が異なることがデータ利活用を複雑化させていると述べた．データはその複製の容易さから，一度インターネット上に情報が流出してしまうと，完全に消去することはほぼ不可能であるため，消費者はデータから個人を特定されることを避けると考えられる．そのため，利用目的が明確でない限り，個人が自身の医療情報やスケジュールを公開することは稀である．すなわち，パーソナルデータ利活用によるサービスのメリットを消費者に提示し，理解と同意が必要となる．個人情報を含んだデータは，プライバシーの観点から他組織との共有は難しい．そのため，企業や個人はデータの公開及び共有を躊躇う傾向にあるのである．介護医療やヘルスケア分野におけるパーソナルビッ

[6] http://www.data-xc.jp/
[7] http://www.kiis.or.jp/OBDF/
[8] https://www.openfogconsortium.org/

グデータ関連研究（Sano & Picard, 2013; 髙玉, 2014 など）では，その需要や可能性に期待が高まっている一方で，個人の識別性やプライバシーの問題が指摘されている．また，データ利活用における個人情報とパーソナルデータの取り扱いと法制度については，国によって状況が異なり，様々な問題が議論されている（石井, 2014; 森, 2014; 高崎, 2014 など）．

さらに，2013 年に JR 東日本が Suica の乗降履歴情報を一般企業に売却した問題は記憶に新しい．データ利活用の機運が高まる中で，個人情報とは何か，パーソナルデータの扱いについて様々な議論を呼んだ．この問題以降，企業はデータの売買に対して一層慎重になったと考えられる．これらの問題に対して，企業のデータ利活用を促進させるために，2015 年初頭に改正個人情報保護法が閣議決定された．何が個人情報に当たるのか，データ利活用のプライバシーへの影響についておよそ 2 年に渡り有識者による検討会が開かれたという．これらの議論から，パーソナルデータに関わる制度改正やビジネスでの情報利活用への影響が議論されるようになった（佐藤, 2015; 中崎, 2015 など）．

同様に，プライバシー保護とセキュリティの観点からデータ管理に多大なコストがかかるという問題も指摘されている．ネットワークにつながっている情報端末は常に情報漏洩のリスクに晒されている．暗号技術の発展に伴い，それらを破る解析技術も同様に発展してきており，コンピューターウィルスや悪質なクラッカーによる顧客情報の抜き取りなどのニュースは絶えない．データの漏洩には多大な損失が伴うため，データ管理及びセキュリティには莫大な費用を投じている組織が多い．

また，オープンガバメント戦略による行政のデータ公開においても，多くの問題が生じていたという事実がある．2014 年 10 月に本格始動した行政のデータカタログサイト DATA.GO.JP は，一度は Web 上に公開されたものの，サイトが閉鎖されるという事態が起こった．当時の記事によれば，公開するデータについてのルール作りが不十分であり，リスクマネジメントの観点から，閉鎖はやむを得なかったと説明されている．

一部の学術界では，オープンデータ，オープンガバメントの高まりから，データの公開及び共有が進んでいる（David, 2003; ヒリナスキエヴィッチ・新谷, 2014 など）．研究においては，データ公開インセンティブがないこと，膨大な研究期間や研究費用をかけて収集したデータを共有することの躊躇などが考えられるが，学術的に価値のあるデータを公開することで研究成果として認める権威あるジャーナルも生命科学分野を中心に登場している．データ保有者は自身のデータを研究成果として認められ，データ利用者及び分析者は，膨大なデータ取得コストをかけることなく，貴重なデータを入手できるところにメリットがある．しかし，あらゆる分野でデータ公開戦略が成立するわけではない．データのプライバシー及び個人の識別性，そしてビジネス機会の損失のリスクなどから，民間企業で蓄積されている膨大なデータを公開することは不可能である．

また，公開され他のデータを結合し，統一的なアクセスを可能にしている LOD においても，実用性の高いインフラとしての機能はまだまだ不十分であると言われている．LOD としてリンクされ一般にアクセス可能なデータは，あくまでも一部の行政や研究機関において公開されているオープンデータに留まっている．そのため，個人，企業，その他の研究機関で収集・蓄積されているオープンになっていない有益なデータの存在を知るための手段がほとんどない．アクセスできないだけでなく，どのようなデータを取得し，どのような意思決定に用いているのかという情報及び知識でさえ入手不可能である．

データの管理コストの他に，データ取得コストも重要な問題の一つである．高度なセンサーを用いて膨大なデータが取得・蓄積されているが，それらを取得・蓄積するのにもコストがかかる．コストに見合った意思決定ができているのか，またはデータの活用による効果を評価できているのか，疑問に思っている機関も多いかもしれない．良質なデータが取得できていても，それらを分析する技術やストレージが不十分であれば，宝の持ち腐れとなってしまうことが懸念される．さらに大企業であれば，一組織内でさえ部署ごとにデータを共有できていないという現状もデータ利活用における問題の一つとして考えられる．例えば，同じ会社内の他の部署が取得しているデータとほぼ同じデータを別の部署が多大なコストを欠けて再取得してしまっていることも考えられる．誰が，どこに，どのような形式でデータを保管しているのか知る方法がなく，同じ組織内でも共有できていない．つまり，データは有効な手段によって効率的に再利用されなければ，多くの無駄を生み出してしまっている可能性があるのである．

以上のように，保有するデータおよび組織や分野を超えた異なる領域のデータから新しい知識を獲得し，新ビジネスの創出や価値の創造という潜在的な可能性への期待は高まっている一方で，データ管理コストやセキュリティ，プライバシー，法制度の観点から様々な問題が指摘されていることが理解できる．そこで，データの公開・共有を強制するのではなく，市場の原理で利用者が必要なデータを選び，所有者と交渉の末に入手することができる，データ市場というイノベーション創出環境におけるデータの価値化プラットフォームが議論され，発展してきた．

2.3 データ市場

2.3.1 データ市場概説

前節で述べたように，オープンデータ，オープンガバメントの高まりから，データの公開戦略が学術界と行政において進んでいる．しかし，あらゆる分野でデータ公開戦略が成立するわけではない．プライバシー問題，価値が未策定なデータの共有による機会の損失可能性，データの用途の不透明性などの潜在リスクから，企業及び個人が蓄積している膨大なデータを公開することは不可能であり，昨今の分野横断的なデータ利活用に対する期待に応えるためには，データ公開という戦略には限界がある．そこで，データ利活用を促進させるため，市場における交換という戦略（Ohsawa et al., 2013; 大澤ら, 2017 など）が有効となると考えられる．

データ市場では，データは商材として市場に提供され，データ保有者，利用者，分析者らがデータ及びその利用方法について理解を深めるコミュニケーションが行われる．このプロセスにより，市場はデータの価値化，交換及び売買の場となる．交換とは，ステークホルダー間で合意された条件を設定してデータを取引する戦略を意味する．2013 年から大澤らは，データ公開のリスクを低減し，データを秘匿にしたまま，CD ショップのジャケットのようにデータの概要情報（データジャケット：Data Jacket）を公開することで，ユーザーがデータの使途を検討し，データ所有者との交渉を経て入手価格などの取得条件を決めるデータ市場を提案している．しかし，データ市場という概念自体は新しいものではない．19 世紀中期に Paul Reuter がパリとロンドン間で電報による株式情報を売買していたという記録が残っている（Dumbill, 2012）．そして 20 世紀中期から，Thomson Reuters 社や Bloomberg 社などがデータアグリゲータ（データを仲介する事業者）として利用者に対し，データの提供をビジネス化してきた．また，1980 年代にはデータ取得製品（センサー）とデータ送信技術の発展に伴い，データ市場が活性化することに言及されている（Electronics and Power, 1980）．その後，計算機の高度化と普及により多様なデータがやり取り可能となると，Web を新たなプラットフォームとするデータ市場が誕生した．すでに，Microsoft Azure[9]，Data Market（2014 年 10 月 Qlik に買収），KDnuggets[10]，Factual[11]，Infochimps[12]といった，データを売買・交換するサービスが登場してきた．

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）」が提案されている（秋山ら, 2013; 砂原ら, 2014）．情報銀行とは，

[9] https://azure.microsoft.com/ja-jp/marketplace/
[10] http://www.kdnuggets.com/
[11] https://www.factual.com/
[12] http://www.infochimps.com/

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．情報銀行のシステムを支える技術として Personal Data Store（PDS）が提案されている．PDS は個人が本人のデータを電子的に蓄積・保管して他者と共有し活用可能とする仕組みを意味する（橋田, 2014）．個人のデータの分散管理によって，集中管理による漏洩リスクを低減し，かつ導入及び運用コストが削減できるとしている．PDS を支える技術として，暗号化，匿名化などの諸技術が発展している．

さらに実業界においては，前節で述べたコンソーシアムなどの興隆に加え，日本データ取引所[13]がデータの取引を仲介する市場を開拓し，EverySense 社[14]が IoT データ流通マーケットプレイスなど展開するように，異なる分野のデータを取引きするためのプラットフォーム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提案され，サービスとして社会実装されてきている．

以上に概観したように，実社会において分野を横断したデータの交換と流通，売買の場である多様な形態のデータの市場が萌芽し始めていることが分かる．

2.3.2 データ市場における問題

データを交換し，取引きするデータ市場の形態が提案され，実社会に展開されているが，それらには様々な問題が存在する．Web を新たなプラットフォームとするデータ市場が出てきているものの，これらの Web サービスは，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十分ではない．なぜなら，Web 上で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は期待できないからである．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．よって，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に関する情報は，市場に登場することができない．そのため，価値あるデータを選んで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．

情報銀行及び PDS は，データ流通環境整備検討会として内閣官房にて議論が行われてい

[13] http://www.j-dex.co.jp/
[14] https://every-sense.com/iot-data/

る．しかし，データ提供者である個人の情報漏えいリスクや自身の情報を信託する不安を解消するためには，まずデータ利活用に関係するステークホルダー間の利活用方法に関する議論が十分行われることが重要であると考えられる．データの提供者は自身のデータの利用方法が分からなければデータの提供をするインセンティブがなく，データの利活用に到達しない．

日本データ取引所や EverySense 社のデータ流通マーケットプレイスなどの実業界においても同様の議論が成立する．本来の市場とは商材の提供者と消費者の間での「提案」と「評価」というコミュニケーションを行うイノベーションの場である．このような市場の原理をデータにも適用し，データ利活用によるイノベーション創出の場が「データ市場」である．すなわち，データ市場は，データを商材と見なし，市場の原理からデータの価値について検討することで，データ保有者（提供者）や消費者（利用者）がデータについての理解を深めるコミュニケーションを活性化させる場である必要がある．

以上に概観したように，実社会において分野を横断したデータの交換と流通，売買の場であるデータの市場が創生し始めているが，データの活用方法に関する知識の蓄積，そしてデータに関わるステークホルダー間のコミュニケーションが実現し，データの価値化とイノベーション創出が有効に働く環境としては十分ではない．そこで，データ市場の創出を支援する手法が必要であると考えられる．

2.4 データ市場創出支援技術の必要性

2.3 節にて述べたように，実社会において分野を横断したデータの交換と流通，売買の場であるデータの市場が誕生しているが，現段階ではデータの価値化とイノベーション創出という市場の機能が有効に働いているとは言い難い．

Purchasing and Supply Management 関連研究では, データ取得コストの節約, 時間の短縮性, 他のデータとの結合可能性による新しい知見の獲得などのメリットがあることから，データの二次利用には潜在的な有用性があると指摘されている（Rabinovich & Cheon, 2011; Ellram & Tate, 2016）．しかし同時に，利用目的や取得意図の異なるデータの二次利用について，分析に適切なデータを検索すること，データを解釈し，結果を理解することに膨大な時間や労力がかかることが課題であるとも述べている．データに関する情報や知識が入手困難である現場では，データ保有者とデータ利用者，そしてデータ分析者などのデータ市場に関わるステークホルダー間のコミュニケーションが十分に行われることは期待できない．すなわち，データ利活用を促進及び組織間のデータ交換においては，まず，世の中にどのようなデータが存在し，誰が保有しているのか，どのように取得されたのかという情報を入手可能な状態にすることが必要であると考えられる．

さらに，Ellram & Tate（2016）は要求に合致した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なければならない部分に重要な問題が存在しているとしている．しかし，Bollier（2010）の指摘では，データの組み合わせから新しい価値を導くという期待はあるものの，異種のデータの組み合わせは客観的な解釈を難しくする可能性がある．また，Boyd & Crawford（2012）はそれぞれのデータの意味を考慮しなければデータの量には意味が無いと述べた．すなわち，ビッグデータ自体ではなく，様々なドメインに蓄積されているスモールデータの価値を理解することの重要性が指摘されている．以上の議論により，データ利活用においては，個々のデータへの理解と適切な組み合わせから導かれる仮説の設定が重要であると言えるだろう．しかし，データの組み合わせは膨大であり，データが増加すれば導ける仮説も指数的に増加するため，取得意図を考慮し，様々な領域に偏在するデータ群のあらゆる組み合わせを考慮することは極めて困難であると言える．すなわち，データの意味や適切な組み合わせを含む利活用方法についてステークホルダー間（データ保有者や利用者）で議論する場を設定する必要がある．さらに，目的に適合したデータの取得及び対象データの様相を十分理解した分析を行うことが重要であるため，議論されたデータ利活用案を実行に移すための事業計画や分析プランであるシナリオ作成を支援することも必要であると考えられる．

また，今日のデータ市場では異なる分野のどのようなデータが結合可能であり，どのような仮説を検証できるのかという知識は共有及び確立されていない．データ市場がデータ

を商材としたイノベーションの場としての機能を有するためには，様々な背景知識を有するステークホルダーの専門知をもって検討されたデータ利活用案や解決可能な問題に関する情報がデータ利活用知識として蓄積され，再利用可能な状態であることが必要であると考えられる．

以上の諸問題を考慮すると，データの利用価値の検討からデータの価値化と交換，流通を促し，データ利活用支援する技術として以下の機能を有する手法が有効であると考えられる．

① 誰がどのような形でデータを保有しているのか知る方法が少ないため，世の中にどのようなデータが存在し，誰が保有しているのか，どのように取得されたのかという情報を入手可能な状態にするための技術及びプラットフォーム．

② ①を用い，単一のデータソースだけではなく，複数のデータの適切な組み合わせやデータの利活用方法についてステークホルダー間で議論可能とする場の設計．

③ 異なる分野のどのようなデータが結合可能であり，どのような仮説を検証できるのか，どのような問題解決が可能なのかというデータ利活用知識の蓄積技術．

データ市場創出を支援するため，本研究では以上の①，②及び③に対応したデータ市場創出支援技術について扱う．

2.5 本章のまとめ

本章では，データの人間社会における役割と位置づけについて述べ，データ市場がどのような社会的要請から発展してきたのかを関連研究及び事例を示しながら明らかにした．

近年，蓄積された膨大なデータを再利用し，様々な分析手法を用いた新しい価値の発見や意思決定に役立てようという動きが活発になってきている．ブログや SNS の流行，ビッグデータ，オープンデータ，IoT などの分野を横断したデータの組合せと利活用によって，既存のサービスの付加価値向上や新製品の開発などのデータ駆動型イノベーションに対する期待が高まってきている．

しかし，データの価値は，様々な背景知識や利用目的を持つステークホルダーによって異なることが認知科学の観点から説明されてきた．そのため，データ利活用とデータに基づく意思決定への期待の高まりがある一方で，データの価値化は難しく，データの交換や利活用については，プライバシーの問題やデータの用途の不透明性などの様々な問題が指摘されてきた．また，異分野のデータ結合による知識発見が期待されているものの，実際は異なる分野のどのようなデータが結合可能であり，どのような仮説を検証できるのかという知識が確立されていないのが現状である．そこで，データの公開・共有を強制するのではなく，市場の原理で利用者が必要なデータを選び，所有者と交渉の末に入手することができる，データ市場というイノベーション創出環境におけるデータの価値化プラットフォームが議論され，発展してきた．

しかしデータを交換し，取引するいくつかのデータ市場の形態が提案され，実社会に展開されているが，それらには様々な問題が存在する．Web を新たなプラットフォームとするデータ市場では，データの表層的な情報を Web 上に陳列するだけのサービスに留まっており，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不十分である．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．

そのため，データの活用方法に関する知識の蓄積，そしてデータに関わるステークホルダー間のコミュニケーションが実現し，データの価値化とイノベーション創出が有効に働く環境としてのデータ市場を整備するための支援手法が必要である．

次章では本章の議論及びデータ市場に関わる課題に対する本研究の着眼点とアプローチについて説明する．

第 3 章　研究の目的 –データ利活用における行動計画支援–

第 2 章では本研究の背景及び課題について先行研究を概観しつつ論じた．本章では，第 2 章でまとめた課題に対する本研究の目的とアプローチについて本研究にて提案する手法の概説を含めて論じる．本章は 4 つの節で構成されている．3.1 節では，本研究が対象とする課題と本研究の新規性について述べる．3.2 節では，データ市場と実験的データ市場について，本研究のアプローチ方法，データの利用価値の観測方法について論じる．そして，3.3 節では，今までの議論を踏まえて設計した本研究の提案手法についての基本的なアイデアを述べる．そして，3.4 節にて，本章のまとめを行う．

3.1 本研究の扱う課題と新規性

データ市場は今まで世の中に出てこなかったデータ，ステークホルダー，知識が登場する新しい市場であることはすでに述べた．第 2 章で論じたように，現在のデータ市場においては，分野を横断したデータの価値を策定するための背景知識が適切な形で蓄積されてこなかった．そのため，既存の分析手法を適用することで分析結果が得られるデータもあるが，その多くはまだ適用する分析手法が定まっておらず，検討する手法も確立していない．すなわち，データの蓄積方法は積極的に議論されてきたが，既存の知識やモデルでは扱えないデータを含むデータ利活用知識の蓄積方法及びその評価方法については十分に議論されてこなかったと言える．

また，第 4 章 4.2 節にて説明するデータ利活用方法検討ワークショップ Innovators Marketplace on Data Jackets（IMDJ）及び第 5 章 5.1 節のアクション・プランニング（AP）にてデータの価値が策定され，データ利活用が促進されることは示されていたが，データ利活用に関わる潜在的なステークホルダーを発見し，当該データの価値を検討するための手法及びプロセスについてはモデル化されてこなかった．そのため，既存の分析手法では扱えない新しいデータが登場した際に，どのように利活用方法を検討するのかというデータの価値化と利活用方法のプロセス，そして，データだけでなく，ステークホルダーやデータに含まれる変数を対象とした手法を研究することは，データ市場の創成及び促進について重要であると考えることができる．

そこで，本研究では，データ市場において，どのようなデータを使うことによりどのような問題が解決可能であるのか，というデータ利活用知識を蓄積し，再利用する仕組みを提案する．そして，データ利活用方法検討時の人間の検索行動，筆記行動，議論プロセスに着目し，データ利活用方法を検討する人間の知識獲得とシナリオ生成を支援する検索システムの実装と，その性能評価を行う．そして，既存の知識やモデルでは扱えないデータの利活用法及び潜在的なステークホルダーの発見が支援されるとともに，データ利活用知

識ベースが更新される．そして，それらを利用して新たに知識を獲得する意思決定者の行動が改善さることを示す．以上に述べた目的及びアプローチによって，データ市場における意思決定者の実行動の計画支援が可能となると考える．上記の課題及び課題に対するアプローチを踏まえ，本研究の新規性について以下に列挙する．

1. データ利活用知識の蓄積とその評価（第 4 章）
   データの概要情報だけでなく,「どのデータの利用がどのような問題解決に役立ち得るのか」という過去に検討されたデータ利活用知識を構造化し，再利用する仕組みを提案し，アプリケーションを開発したこと．また，実験的データ市場において価値を認めら得るデータの特徴について実験的に考察したこと．

2. データ利活用における行動計画支援のためのシナリオの構造化（第 5 章）
   第 4 章におけるデータ利活用方法検討の支援だけでなく，実社会とのインタラクションを含めた実行動シナリオの構造化と再利用方法を議論したこと．そして，シナリオ生成における意思決定者の行動の特徴について実験的に考察したこと．さらに，データだけでなく，データに含まれる変数の特徴をモデル化し，自然言語による変数名の検索アプリケーションを開発し，実験的に評価したこと．

3. データ市場活性化のための支援ツールの開発（第 4 章，第 5 章）
   データの概要情報だけでなく，ステークホルダー，変数名，リソースなどの関連要素の表出を支援するアプリケーションの開発と評価を行ったこと．特に，データ概要情報の検索システム Data Jacket Store（第 4 章），ステークホルダー推薦システム Resource Finder（第 5 章）変数名の推定システム VARUABLE QUEST（第 5 章）について実験を行い，アプリケーションの有効性を評価した．

4. 実装的データ市場における知識獲得のプロセスの観察と応用実験（第 6 章）
   実験的データ市場だけでなく，実社会とのインタラクションを含めた実装的データ市場において，データ利活用方法の検討のみならず，データ取得，実分析，フィードバックまでの一連のプロセスにおける意思決定者の行動と課題発見のプロセスを議論したこと．また，本研究の提案手法によって，事業者の新しいデータの発見や異なる事業者とのインタラクションによって新規事業創出が促進されたことが報告されており，データ市場の創出と発展，そして DDI に貢献したこと．

3.2 データ市場と実験的データ市場

3.2.1 データ市場の観察

データ市場では，データは商材として市場に提供され，データ保有者，利用者，分析者らがデータ及びその利用方法について理解し合いながら，データの価値化による交換及び売買が行われる．株式市場や金融市場のように取引の記録が追跡可能であれば，既知の手法を用いてモデル化することが可能かもしれないが，実際のデータ市場におけるデータの取引やステークホルダーの目的や意図を直接観測することは難しい．人工市場（和泉, 2003 など）のように，すでに確立された経済理論や金融モデルを用いて市場のメカニズムを解明する方法が考えられるが，本研究で扱うデータ市場は，従来の市場に出てこなかった新しいデータ，ステークホルダー，知識が登場する市場であり，予めすべての要素を考慮したモデルを構築することは適切ではない．また，データは情報や知識と比較し，既存の経済財の価格決定メカニズムが適用可能とは示されておらず，解明されていない．

以上の理由から，本研究ではデータ市場を模した実験的データ市場をワークショップ形式で実現することで，実際のデータ市場のデザインにおいて重要な知見を得るというアプローチを行う．実際のデータ市場に参画するステークホルダー（実業家，研究者，データ分析者，市民など）を参加者とし，データ市場における各ステークホルダーの意図や目的が反映される場を設計する．すなわち本研究では，データ利活用方法検討ワークショップ Innovators Marketplace on Data Jackets（IMDJ）及びアクション・プランニングを実験的なデータ市場として設計し，ワークショップにおける参加者の行動を観察することでデータ市場を理解しモデル化することを試みる．実験的データ市場は，実社会にあるデータ市場のモデルである．ここでいうモデルとは，実社会の複雑な現象を分析する時に，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したものを意味する．実社会のデータ市場を模したワークショップを提案しているのは，実際の複雑なデータ市場において，ステークホルダーが行っている学習や行動，価値づけがどのような相互作用によって市場の特徴を形成しているのか理解するためである．つまり，実験的データ市場におけるステークホルダーの挙動を観察することによって，現実の市場現象の分析，現場の支援ツールの提案，既存の市場理論の検証ができる．

行動経済学では，条件が統制された実験室内で市場を再現し，人間が参加する実験を行う（Friedman & Sunder, 1994; Kagel & Roth, 1995 など）．実験的データ市場では，参加者に架空の資産を与え，様々な条件下でデータの利活用方法や要求を提示し，データの交換及び利活用方法を検討してもらう．本研究のアプローチは，この行動経済学における実験市場の手法を用いて実験を行う点で共通している．

3.2.2 データの利用価値と利用期待度

2.1 節では，データは文字，数値，記号として記述可能な事実であり，情報はデータに文脈が与えられたものであると定義した．言い換えれば，データとは的確な利用文脈が発見されていないものであると言える．つまり，データは市場においてデータとは未だ価値が見出されていない材を意味する．

Boisot & Canals（2004）の指摘した 2000 年代初頭におけるデータマイニング分野では，人々はデータにお金を払うのではなく，データから得られた知識にお金を払う傾向があったという．その理由は，データ分析や情報抽出は専門性が高くかつ高度な技術が必要だからであった. そのため, データは価値の評価が難しく, 従来の経済交流（economic exchange）の仕組みに取り入れるのは難しいため，データは経済財として見なすことができるものの，価格設定が困難であるという指摘をしていた．すなわち，知識や情報は実社会に働きかける行動（actions）に結びつくため，価値を決定しやすく，経済財として価格設定が比較的容易である一方で，価値ある情報が未だ見出されていないデータは経済財として価値化することが困難な商材であると言える．

それでは，データの利用価値はどのように評価されるべきであろうか．データは文脈が与えられることによって，人間の意思決定に必要な情報となり，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．データに文脈を与える行為とは，すなわち分析である．分析とは，「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」であると一般的に認識されている．従来の分析では，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，データを得ることで十分に物事の内容や性質を明らかにすることが可能であったかもしれない．しかし，取得可能なデータが多岐に渡り，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，異なる領域の様々な知識を転用，あるいは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てきたと考えられる．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をも包含している．これらは，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であるとする Ellram & Tate（2016）の主張とも合致する．

しかし，事前にデータを入手し，分析することでデータに文脈を与えることは困難である. 2.2 節にて議論したように, 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からである．そのため，分析によって利用価値を評価する前にデータ自体を入手することが一般的に困難である．そこで，データを

入手する前にデータの利用価値を策定する必要がある．データ市場において，データとは未だ価値が見出されていない商材であることを考慮すると，データの価値を直接測定するのではなく，データにどの程度利用価値があるのか期待する度合い（利用期待度）を測定することが適当であると考えることができる．すなわち，データ自体を分析する以前に，データについてのデータであるメタデータを用いてデータの利用期待度を評価することでデータの価値を策定するというアプローチを行う．

また，個人のデータ理解によって解が促進し，データの価値を決定することができるかもしれないが，データから価値を策定するプロセスは単一のエージェントではなく，多様な目的と意思を持った複数のエージェントによる相互作用で成立する．「三人寄れば文殊の知恵」という故事のように，エージェント間でのコミュニケーションによって，当該データの新しい使い方や従来と異なる問題解決方法への転用などが期待できる．特にデータ市場において重要なのは，データ市場に関わる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付けと評価，すなわち利用期待度である．利用期待度によってデータの価値が定まれば，データの交換・売買という経済交流が発生し，データの市場が形成される．売買が促進されれば，データ保有者の重要なデータの公開や販売の動機にもつながり，データの市場が活性化する．データを経済財として価値化するためには，複数のステークホルダーの相互作用によるデータ利活用方法の発見と評価が重要であると考えられる．

3.3 提案手法概説

3.3.1 データ利活用知識及びシナリオの構造化

データ市場は今まで世の中に出てこなかったデータ，ステークホルダー，知識が登場する新しい市場であるが，昨今のデータ市場では，データの価値を策定するための背景知識が適切な形で蓄積されてこなかった．また，データの蓄積方法は積極的に議論されてきたが，既存の知識やモデルでは扱えないデータを含むデータ利活用知識の蓄積方法及びその評価方法については十分に議論されてこなかった．

本研究では，データ市場における DDI に貢献することを目的とし，データ利活用知識構造化と検索システムによる人間のデータ利活用シナリオ生成支援手法の提案を行う．データの概要情報であるデータジャケット（第 4 章 4.1 節）だけでなく，過去のデータ利活用方法検討ワークショップ IMDJ（第 4 章 4.2 節）において議論された要求，データ利活用案によって価値を認められたデータの関係をデータ利活用知識としてモデル化する．データ利活用知識構造化により，ユーザーが自分と異なる視点を持つ過去のユーザーが考案したデータの使い道を発見したり，過去の別の人が考案したデータ結合案に注目することによって役に立つデータジャケットを探し出すことが可能となる．

さらに，第 4 章のデータ利活用知識の構造化を拡張し，人間の意思決定を支援するシナリオ創出手法アクション・プランニング（第 5 章 5.1 節）によって生成されたシナリオの構造化と再利用する仕組みを提案する．過去のワークショップで検討されたデータ利活用シナリオに含まれていた知識を再利用し，シナリオ実現に欠けている知識を補完することが可能となる．シナリオを介して潜在的なビジネスパートナーを推定・推薦するだけでなく，敵対するステークホルダーについても事前に予期できるため，シナリオを考案する時点で対策を検討することが可能となり，実行動におけるリスクを低減できると考えられる．

そして，新たにデータを取得する意思決定者の支援手法として，データの説明文であるデータ概要から，そのデータに含まれるであろう変数ラベルを推定する手法を提案する．変数ラベルの類似度及び変数ラベルの共起性を考慮することで，変数ラベルが未知のデータ概要からそのデータに含まれる可能性の高い変数ラベルを推定できることを示す．

データ市場において，データに文脈を与え，実社会において有用である事業とするには，データの組合せだけでなく，ステークホルダーやリソースといったデータに関わる諸要素の関連性を考慮した行動計画の検討が重要であるため，「データ」，「ソリューション」，「要求」だけでなく，「ステークホルダー」，「変数ラベル」を含めた要素を構造化したシナリオ及びデータ概要情報の再利用は新たな知識獲得に有用であることを示す．

3.3.2 本研究のデータ駆動型イノベーションへの貢献

市場における交換という戦略によって分野を横断したデータ駆動型イノベーションを推進しようとする様々な形態のデータ市場が創生されてきている．しかし，Web を新たなプラットフォームとするデータ市場では，データの表層的な情報を Web 上に陳列するだけのサービスに留まっており，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十分ではない．また，膨大なデータから必要な知識を発見することが困難であり，データ市場において複数の領域にまたがって存在するデータ，ステークホルダー，ツールなどすべての要素を考慮することは難しい．それ故，意思決定者の異なる価値観や関心を持つ多様な背景知識，意図に対応して適切に設計し，構造化された知識ベースとそれを検索するシステムが必要となる．

本提案手法及び本研究の実験で得られた知見は，データ市場はオープンデータに代表される公開可能データのみに閉じられた場ではなく，公開が難しい個人や企業のデータ及びその保有者を巻き込むイノベーションの場として機能し得ることを示すと同時に，データの価値化とイノベーション創出が有効に働く環境としてデータ市場を整備するための支援手法として機能することが期待できる．

3.4 本章のまとめ

本章では,第 2 章でまとめた課題に対する本研究の着眼点とアプローチについて論じた．先行研究では，データの蓄積方法については積極的に議論されてきたが，既存の知識やモデルでは扱えないデータを含むデータ利活用知識の蓄積方法及びその評価方法については十分に議論されてこなかった．

3.1 節では，データ市場におけるデータ利活用知識を蓄積し，再利用する仕組みを提案することを課題とともに明示し，次章以降における議論の概要を述べた．また，本研究の課題に対するアプローチを踏まえ，本研究の新規性についてまとめた．

3.2 節では，データ市場と実験的データ市場について，本研究のアプローチ方法及びデータの利用価値の観測方法について論じた．実際のデータ市場に参画するステークホルダーを参加者とし，データ市場における各ステークホルダーの意図や目的が反映されるデータ市場を模した実験的データ市場をワークショップ形式（データ利活用方法検討ワークショップ IMDJ 及びアクション・プランニング）で設計する．実業家，研究者，データ分析者，社会で特定の役割を持つ個人などのデータ市場におけるステークホルダーを参加者とし，ワークショップにおける参加者の行動を観察し，データ市場の仕組みを理解しモデル化することを試みる．実験的データ市場を観察することで，実際のデータ市場のデザインにおいて重要な知見を得るという手法について述べた．

3.3 節では，上記の課題とアプローチを踏まえ，本研究の提案手法についての基本的なアイデアと概要について説明し，本研究のデータ駆動型イノベーションへの貢献について明示した．

第 4 章　データ利活用知識の構造化と検索システム

データ市場において，データの蓄積方法については積極的に議論されてきたが，既存の知識やモデルでは扱えないデータを含むデータ利活用知識の蓄積方法及びその評価方法については十分に議論されてこなかったことを説明してきた．本章では，データ市場における知識の蓄積のためのデータ利活用知識の構造化と検索システムについて論じる．

4.1 節では本研究の核となる基礎技術であるデータジャケットについて概説する．そして，4.2 節ではデータジャケットを用いたデータ利活用方法検討ワークショップ Innovators Marketplace on Data Jackets（IMDJ）について説明する．4.3 節では，ユーザーのデータ利活用方法検討を支援するためのデータ記述モデルとデータジャケットの構造化について検討する．4.4 節では，ユーザーのデータ利活用方法検討を支援するためのデータ利活用知識の構造化について議論し，データ概要情報検索システム Data Jacket Store（DJ ストア）を実装，その性能を評価する．また，4.5 節では価値が認められるデータの特徴について，データの共有条件に着目し，実験的データ市場における利用回数や DJ ストアの閲覧回数と比較し，考察を行う．そして，4.6 節には本章のまとめを行う．

4.1 データジャケット（Data Jacket）

前述したように，分野を横断したデータ利活用に対する期待が高まっているものの，データ利用者が自身の興味・関心のあるデータを誰が保有しているのかということを知る方法は少ない．また，データ保有者が自身の保有するデータをビジネス機会の損失リスクを低減して市場に提供することは難しい．そこで，大澤らは，データ自体は秘匿にしたまま，データに関する情報を共有するための技術として，データジャケットを提案している（Ohsawa et al., 2013; 大澤, 2014a）．データジャケットとは，あるデータがどのような情報を有しているのかを説明するための概要情報，すなわちメタデータである．データジャケットは CD や DVD は購入しないと中身の音楽や映像は閲覧することはできないという性質から着想を得ている．CD や DVD のジャケットにはアーティスト名や映画の出演者，コンテンツの長さなどのコンテンツに関する説明文が記述されている．我々はジャケットに記述された説明文を読むことにより，中身であるコンテンツについて理解し，その価値を考えることができる．同様に，データ本体は購入などにより入手しなければ閲覧することはできないが，ジャケットに書かれた中身に関する説明を読むことにより，データの中身や価値について理解することが可能となる．データ保有者は，データジャケットによってデータ本体を公開することなく，データに関する情報をデータ利用者に理解してもらい，利用価値を検討してもらうことが可能になる．

4.1.1 データジャケット概説

データジャケット（Data Jacket : DJ）はデータの中身ではなく，データの概要情報（データ内の変数名，保存形式，収集方法など）を共有し，データの価値を検討可能にする方法である．個人を識別する情報を含む共有不可能なデータでも，DJ にすることでセキュリティ上のリスクを低減させてデータに関する情報が共有可能となる．例えば，商品の購買履歴データには氏名，性別，支払金額などの個人を識別する情報が含まれる．そのため，これらのデータを Web 上に公開して共有可能な状態にすることは通常は困難である．しかし，購買履歴データを「氏名」,「性別」,「支払金額」といった公開可能な変数名（変数ラベル）としてメタデータ化すれば，個人を識別する情報は秘匿のまま，データに関する情報が共有可能となる．

また，変数ラベルだけでなく，DJ ではタイトル，データ概要，データの収集方法，保存形式，共有条件など 12 の記述項目を定義している．これらの DJ に記述された内容を読めば，例えば，そのデータがどのような形式のもので，誰がどのような意図で取得したのか，またデータ取得にかかったコストや期待する分析成果などを理解することが可能となる．さらに，DJ として記述されたデータに関する構造化された情報に対してテキストマイニングのツールなどを適用することにより，人間だけでなく計算機においても可読となり，データの関連性を可視化することも可能となる．データ同士の繋がりを共有条件や変数ラベルから理解することができ，仮説の生成などを支援することが可能となる．

本研究では DJ の入力フォームを DJ Site（DJ サイト）[15]に開設し，ビジネスパーソン，専門家，データサイエンティストなどから広く DJ を収集した．データ保有者は公開可能な部分のみを記入し，DJ を作成する．図 4-1 は DJ 入力フォームから登録された DJ の一例である．

[15] https://sites.google.com/site/datajackets/

DJ No. XX【東京都の街路灯管理データ】	
概要	東京都が管理する都道における街路灯の設置・管理に関するデータ
収集方法・コスト	街路灯を設置した会社との共同作業によるデータを取得。入手には都庁の事務所に相談が必要。
共有条件	条件・交渉により共有可
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	・地図へのマッピング ・地域あたりの明るさを算出
変数ラベル	照明種別、管理番号、柱種、適合ランプ、町名、町名、番地、号、緯度、経度、光束、ランプ等級、器具電力、ランプ電力、色温度、全光束、演色性、定格寿命
分析結果	定格寿命を考慮した効率的なメンテナンスとランプ交換タイミングをアラート

図 4-1　DJ の記入例

4.1.2 データジャケットの記述項目

データ市場において，データを提供するステークホルダーは，データ保有者である．データ保有者は，個人あるいは法人などの組織を意味する．すなわち，データを所有するあらゆる個人・企業・研究機関がデータ保有者になり得る．データ市場における主要プレイヤーの一人であるデータ保有者は，自身が所持するデータを市場に提供する役割を持っている．しかし，前述したように，データ本体の公開はリスクが高い．そのため，自身が保有するデータを DJ として提供することが，データの価値策定と取引の場であるデータ市場において重要となる．

DJ はデータを保有者が自身の意思で記述することを原則としている．また，国や各自治体の統計データや，個人が公開している研究データなど，すでに Web などに一般公開されているデータについては，一般の参加者が入力してもよい．その際はデータの所在などを URL で明示する必要がある．DJ は 2016 年 11 月現在，自然言語による 12 の記述項目を設定し，収集されている．自然言語は多義性の問題を有しており，事物の概念を一意に解釈することを困難にしている要因となる可能性があるという指摘があるかもしれない．しかし，データ市場は，研究者やデータサイエンティストだけのものではなく，データに関わる全てのステークホルダーに対して開かれた市場である．データ市場がデータに関して様々なステークホルダーがコミュニケーションを行う場であるためには，市場に出回るデータに関する情報は，万人が理解できる形式で記述されることが望ましい．KDnuggets など

のデータポータルサイトや The Humanitarian Data Exchange[16]，DATA.GOV.UK[17]などの既存の行政のオープンデータポータルでは，データに関する情報は自然言語にて表現されている．そのため，DJ の記述でも同様に，データについて理解可能にするために自然言語による記入を許容している．

以下，DJ の記述項目の詳細について説明する．

4.1.2.1 データのタイトル

データには名前が付与されている．データのタイトルとは，データの存在を一意に決定するラベルに相当する．DJ として登録されているタイトルには，例えば図 4-1 の例のように，「東京都の街路灯管理データ」であったり，「金融システムレポート」，「火災実験データベース」などがあり得る．しかし，タイトルだけでは，人間の同姓同名のようにデータを一意に定めることができない場合がある．そこで，DJ ではデータのタイトル以外にデータ概要や含まれる変数ラベルなどを収集している．

4.1.2.2 データ概要

データ概要とは，データについて説明するための文章を表す．タイトルだけでは表せないデータを特徴づける説明文となる. DJ 登録の記述ルールにおいては必須条件ではないが，データがどのような情報を含んでいるのかを説明し，データ保有者が自身のデータをデータ利用者やデータ分析者などのデータ市場の他のプレイヤーに理解し，利用方法を考案してもらうために重要な情報となる．

4.1.2.3 データの所有者とその所在

データを保有する企業，機関あるいは個人に関する情報が記載される．すでに行政のサイトなどで公開されているデータについては，その URL などがあることが望ましい．DJ サイトでは所在に関する情報を収集しているが，現在サイト内では公開されていない．

4.1.2.4 データ収集方法やコスト

データをどのように収集したのか，どの程度金銭的コストが生じたのか，といった情報を記載する．データの収集した時期や条件などの詳細情報が載っていれば，データ利用者や分析者はデータ取得者の意図を解して適切な分析方法を提案できるようになる．

[16] https://data.humdata.org/
[17] https://data.gov.uk/

4.1.2.5 データの共有条件

DJ はデータに関する情報を共有するための技術であるが，その背後には実データが存在する．実データは購入により共有が可能であったり，範囲を限定して共有が可能，あるいはまったく共有ができないものなどが存在する．このような共有に関する条件を記入するのが共有条件である．特にオープンデータに代表される共有可能データは「一般に共有してよい」と書かれ，個人や企業の秘匿データはそれぞれの共有条件に合った内容が記述される．

4.1.2.6 データの種類

データに含まれる変数の値の種類を記入する部分である．データには数値データだけでなく，アンケートなどの文字データ，画像，音声などの種類が存在する．この項目では，そのようなデータの特徴に関する情報を記入する．

4.1.2.7 データの保存形式

データが保存されているフォーマットを記入する．画像データであれば JPEG，PNG，ビットマップ形式など様々なフォーマットが存在する．フォーマットによっては，適切な分析手法が適用できない場合があり，その際には適切な変換処理を行う必要がある．フォーマットの情報によって，期待する分析結果を得るためのデータの構造変換方法や適用する処理を議論することが可能となる．

4.1.2.8 変数ラベル

変数ラベルとは，データ固有の変数に関して自然言語によって記述された説明文を意味する．データの結合可能性を議論する際に，変数ラベルを用いて可視化，あるいは仮説を立てる．データ自体に含まれる変数を秘匿のまま，変数ラベルの組み合わせを議論することで大まかな分析シナリオを立てることが可能となる．

4.1.2.9 分析・シミュレーションプロセス

データ分析とは，データに含まれる変数の値を入力とし，あるルールに従って組み合わせ，変換することによって，出力結果を得るプロセスである．ここでは，DJ として登録するデータがどのような理論によって取得されたのか，あるいはどのような分析プロセスを支えるために取得されたデータであるのか，具体的な理論とともに説明する．

4.1.2.10 分析・シミュレーションプロセスの結果

データ分析はデータ内の変数の値に対して分析ツールを適用し，値を変換して可視化などの分析結果を得るプロセスである．ここでは，DJ として登録するデータがどのような分析結果を期待して取得したのかという情報を記入する部分となる．

4.1.2.11 分析・シミュレーションプロセス以外に期待する分析

前項目 4.1.2.10 にて，データから導かれるデータ分析の結果について記入したが，ここでは，従来の利用意図以外の期待する分析方法や転用する分野についての情報を記入する．

4.1.2.12 自由記述（データに関する補足事項）

自由記述の項目では，上記の項目に該当しないデータに関する補足事項を記入する．例えば，データが取得された背景にある社会的問題についての企業理念や，データを用いて分析が行われた事例の紹介（論文，ビジネス事例）などが載せられる．

4.1.3 データ概要情報の先行研究

データ本体ではなく，データのデータであるメタデータからデータ同士の結合を行う試みはいくつか行われてきた．初期のメタデータは図書館のカードベースの索引であると言われている（是津, 2007 など）．図書館カードには著者名，タイトル，本の概要などが記述されており，本がどの棚にあるのかを示す機能があり，カードボックスに収蔵されていた．これらが電子化され，オンライン上で閲覧可能となったものが Online Public Access Catalog（OPAC）である．さらに，Web 上のリソースやメタデータを効率的に検索可能とするための共通語彙として Dublin Core などが提案されてきた．

その後，データ本体ではなく概要情報，すなわちメタデータからデータにアクセス可能にする技術がセマンティックウェブの諸技術が開発されてきた．例えば，行政が二次利用を許可した形で公開するオープンデータを RDF という記述言語で表現することで，メタデータを介して各データベースに含まれるデータに統一的にアクセスする環境の構築を目指す LOD 関連研究が進んでいる（武田, 2011; 大向, 2013; 岡嶋ら, 2015 など）．政府統計の総合窓口 e-Stat[18]などのポータルサイトで有名な統計データとメタデータ交換関連研究の分野では，データに統一的にアクセスするための XML ベースの情報モデルを用いて，様々なデータに含まれる変数の組み合わせから統計解析を容易にするための API やメタデータ記述方法を提供している．以上のように，膨大なデータを相互運用可能な形で取り扱い可能に

[18] http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

する技術として，データの中身やデータの意味を計算機にも理解可能にするセマンティクスが期待され，研究が進んできている．しかし，これらの研究はデータが公開あるいは共有可能であることが条件となっており，異分野のデータ統合を前提とした議論に留まっている．本研究で扱う DJ などのデータ利活用方法検討支援技術は，データが公開されていない状況における異なる領域間のデータ共有及び交換の促進，そして新たなデータ取得支援について言及している点で従来研究とは異なっている．

また，LOD の分野では，変数の名前を述語（predicate）として，変数ごとに統一した語彙の利用を推奨している．しかし，実際に行政が公開しているオープンデータのほとんどは変数の名前に自然言語を用いており，語彙の統一は図られていない．例えば，変数の名前である「住所」と「所在地」,「人口」と「人数」は同じ意味であるが異なる語彙で表現されている．厳密に定義した語彙をデータに適用するには，データ取得者の背景知識及び取得意図を考慮する必要があり，これらの作業には膨大なコストと人手がかかることは容易に想像できる．このような問題がある中で，schema.org[19]のプロジェクトでは，Google，Yahoo!，Microsoft が web の改善を目的として，構造化データの仕様を策定する取り組みを推進している．schema.org におけるスキーマとは，HTML タグを表しており，世の中に存在するあらゆる Web ページ内の情報を構造化するためのタグを提供することを目標としている．例えば,「大学」であれば「組織（EducationalOrganization）」に含まれていることが分かる．このようなタグを HTML などのマークアップ言語に入れ込むことで Web ページ内の内容を構造化することで計算機の可読性を高め，検索エンジンの機能高度化を目指している．しかし，日々生み出される膨大な Web ページやデータすべての統一した語彙を設定することはほとんど不可能である．さらに，世の中にある全ての事象やデータを分類することも極めて困難である．また定量的なデータだけでなく，アンケートにおける自由回答のように，世の中には定性的なデータも数多く存在しており，定義されている語彙数以上の多様な種類の変数が世の中には存在している．つまり，世の中にある全てのデータに統一した語彙を付与することは難しい．たとえ行政のオープンデータで変数名の語彙統一が可能であったとしても，企業や個人の秘匿された全てのデータに対して適用可能にはならないと考えられる．そこで，データジャケットでは，変数の意味を表す変数ラベルはデータ取得者の意図や思想が大きく反映されるため，変数ラベルは統一不可能であるという前提に従って設計されている．つまり，人間がデータに含まれる変数の意味を理解可能とするため，データ固有の変数に関しての説明を自然言語によって記述することを許可している．

また，LOD，schema.org，Dublin Core などの関連研究の一部は国際標準化し，社会実装が

[19] http://schema.org/

進んでいるものの，これらの研究ではメタデータの標準化に偏重するあまり，形式レベルでの議論に留まっているのが現状である．つまり，形式に合わないデータについては議論の対象とされていない．前述のように，統計学の枠組みで表現可能な定量的データ以上に世の中には多様なデータが存在しており，データ市場はそのような今までは市場に登場しなかった新しい多種多様なデータが登場し，扱われるイノベーションの場である．以上の議論を踏まえると，多様なデータの存在を許容し，それらの活用方法について積極的に議論する場が必要であると言えるだろう．その中で，データジャケットは多様なステークホルダーが共存するデータ市場において，人間が認知し，表現し，理解し得る自然言語でデータの概要情報を記述することでデータに関する情報を共有できるフレームワークを提案しているところが，従来研究と異なっている点である．

4.2 Innovators Marketplace on Data Jackets（IMDJ）

4.2.1 IMDJ 概説

実験的データ市場として，データ利活用に対する人間の創造性とデータの価値発見を支援するために，様々なステークホルダーの立場から議論し，解を導くワークショップ手法 Innovators Marketplace on Data Jackets（IMDJ）が提案されている（Ohsawa et al., 2013; 大澤ら, 2017）．IMDJ はデータの概要情報である DJ を入力として，データに関わるステークホルダー間のコミュニケーションから，データ利活用方法の検討が行われる場として機能する．IMDJ では，データ市場に関わり，データの利用方法について議論するプレイヤーの最小単位としてデータの「保有者」,「利用（消費）者」,「提案者」の 3 者を設定している．保有者はデータを保有するプレイヤー，利用者はデータを活用したいと考えているプレイヤー，提案者はデータ利活用方法を提案するプレイヤーである．ワークショップ中，参加者は利用者，提案者，データ保有者を兼任し，保有者は自身の保有するデータに関する情報を DJ として提供する．データ利用者の立場からは，自身が意識している問題を提起し，他の参加者に要求を出す．続いて，提案者の立場から，利用者の要求を深掘りしながら DJ に記述された情報を読み解き，データ利活用案（ソリューション）を提案する．これらのやりとりにより，データ保有者は自身のデータを公開することなく活用方法を知ることができ，利活用価値から利用者と取引に関する交渉を開始させることが期待できる．

IMDJ は計算機と人間の協創プロセスに基づき設計されている．まず，DJ を適切に設計された可視化手法によって可視化する．この可視化によって，潜在的な DJ 同士の繋がりを参加者に見せることで，組合せのパターンを減らし，人間の認知的負荷を低減させる．可視化された DJ 間の潜在的な繋がりから人間が価値を導き出すことで実現に足るデータ利活用と意思決定を促すことができると考えられている（図 4-2）．IMDJ には固定されたルールがあるわけではなく，利用目的や実施形態に合わせて人数や時間などは柔軟に変更してよいと定めている．

また，IMDJ によるデータ利活用案を精緻化し，シナリオを生成するプロセスとして，アクション・プランニング（Hayashi & Ohsawa, 2013 など）が提案されている（詳細は 5.1 節で説明）．IMDJ 及びアクション・プランニングによるデータ利活用シナリオにより，データの利用価値や市場性評価が可能となり，データの売買や共有を行うための市場形成が期待できる．IMDJ 及びアクション・プランニングは，企業の新ビジネス策定ワークショップに用いられ，2014 年度及び 2015 年度には経済産業省主催のデータ駆動型イノベーション創出戦略協議会の調査事業ワークショップ（経済産業省, 2014b; 経済産業省, 2015; 経済産業省, 2016）や，国土交通省のビッグデータ活用のための調査研究（国土交通省, 2016），さらにデータエクスチェンジ・コンソーシアムにおける企業間のデータ連携支援の方法論と標

準化に活用されている（DXC, 2014b）.

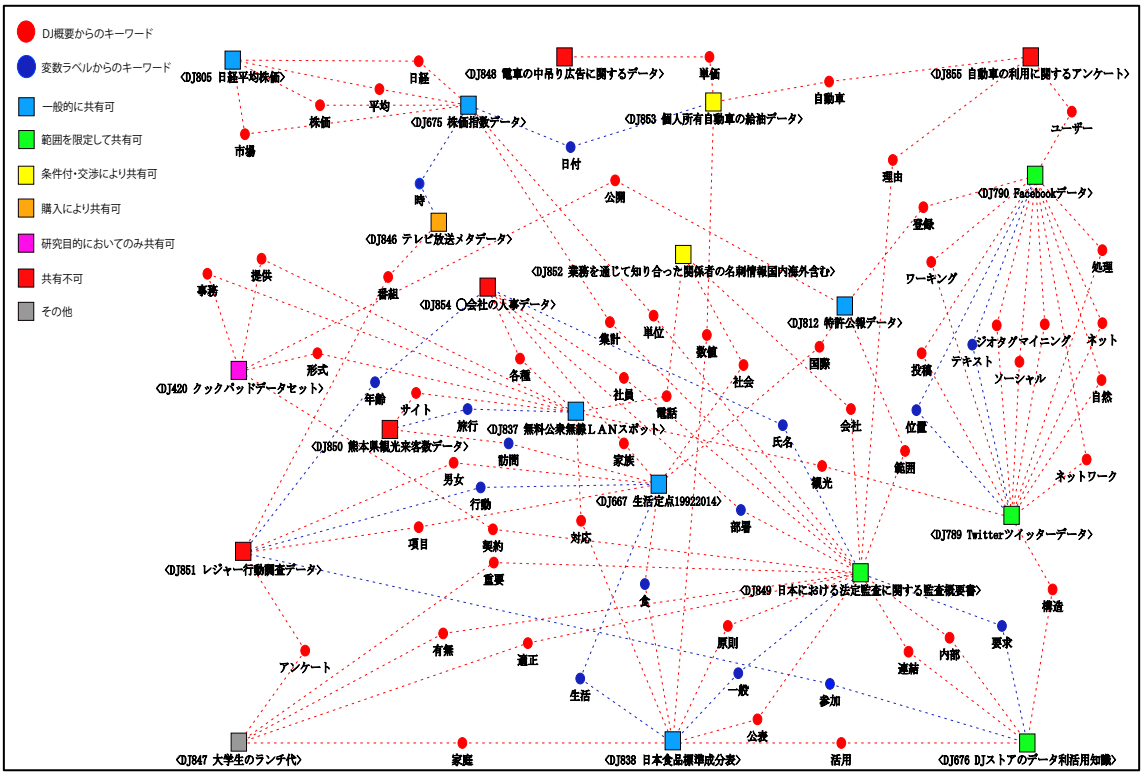

図 4-2　IMDJ で用いる DJ の可視化図の例

（可視化に KeyGraph（Ohsawa et al., 1998）を利用し，一部加工した）

4.2.2 IMDJ におけるデータの価値化プロセス

3.2 節にて，データから価値を策定するプロセスは単一のエージェントではなく，多様な目的と意思を持った複数のエージェントによる相互作用で成立することを述べた．IMDJ よって，データ市場に関わるステークホルダー間のコミュニケーションによる当該データの新しい使い方や従来と異なる問題解決方法への転用などが生起し，データの価値付けと評価（利用期待度）が行われる．データの利用方法が検討されれば，データに文脈が与えられ，データの価値が定まる．データの価値が定まれば，データの交換・売買が行われ，データの市場が起こる．売買が促進されれば，データ保有者の重要なデータの公開や販売の動機にもつながるのである．つまり，データの市場が活性化するのである．IMDJ のプロセスによって，Web 上などにデータの概要情報を陳列するだけでは起こり得ない，ステークホルダー間のコミュニケーションが発生することが期待できる．

適切にデザインされた方法によって集団でのコミュニケーションは創造的になり，アウトプットの質を向上させることがいくつかの従来研究で示されてきた．例えば，IMDJ の前

身である組み合わせアイデア発想手法 Innovators Marketplace（Ohsawa & Nishihara, 2012）においては，ワークショップ参加者同士の批判を奨励することで，ワークショップの質が向上すると述べている．ここでいう批判とは，単純な否定ではなく，相手の意見の問題点を指摘し，反証したり，問題提起を行う発言のことを意味している．このような建設的な批判に加え，IMDJ では「どうして（how)」,「なぜ（why)」といった互いに欠けている視点や知識を補い合うような質問を積極的に用いることを勧めており，同調性を減らすようなコミュニケーションを実践している．特に IMDJ では，データの取引条件（価格や活用方法を発見した暁には共有するなどの約束）を決める交渉も進めながら，データを結合する解析手法とその用途に対する利用者からの批判や要求，評価といったコミュニケーションを通して当該データの異なる利用方法を発見したり，解決すべき問題を洗練する効果があると考えられる．

4.3 データジャケットの収集と構造化

4.3.1 データジャケットの収集方法

筆者は 2013 年 9 月から google フォームを用いてデータジャケット（DJ）の Web 上での収集を開始した．その後，2014 年 5 月から google site を用いて DJ サイトを開設し，ビジネスパーソン，各分野の専門家，研究者，データサイエンティストなどから広く DJ を収集した．データ保有者は DJ 登録者として，データに関する情報のうち，公開可能な部分のみを記入し，DJ サイトに DJ を登録する．図 4-3 は 2013 年 9 月から 2016 年 10 月までの DJ の累積登録件数の推移を表す．企業や組織のデータ利活用への期待は高く，登録されているデータジャケットは増加傾向にあることが分かる．2016 年 11 月現在，1,092 件の DJ が DJ サイトに登録されている[20]．

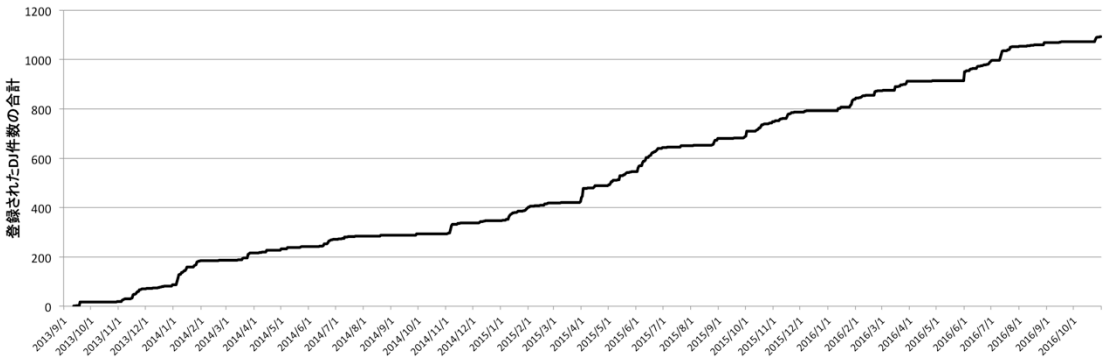

図 4-3　2013 年 9 月から 2016 年 10 月までの DJ 登録件数の累積

4.3.2 データジャケットに含まれるデータの特徴

DJ として登録されたデータがどのようなものであるのかを説明する．なお，いずれも，2016 年 11 月までに収集された DJ を対象としている．また，DJ はデータ保有者が自身の保有するデータについての情報で公開可能な範囲で記述するため，全ての登録 DJ が 12 件の DJ の記述項目を満たしているわけではない．また，データによっては複数のフォーマット（CSV と RDF など）やデータの種類（テキストと数値の組合せなど）を有している．そのため，以下の図の DJ 件数は重複を含んでいる．

図 4-4 は DJ に登録されたデータの保存形式を表す．最も多いのは，CSV あるいは XLS の表形式であり，約半数の 45%を占めている．続いて多いのは，txt 形式で保存されている非構造化データである（約 20%）．CSV 及び txt 形式はオープンデータとして行政などの機関で公開しているデータのフォーマットに多用されている．なぜなら，行政のオープンデ

[20] この件数は DJ サイトで収集されたものだけを含んだ数となる．DJ は DJ サイトで収集されているものだけではなく，企業におけるインハウスのワークショップでは個別に DJ を集めることもある．

ータなど二次利用可能なデータは一般的に汎用性が高く，加工しやすい形式で保存されている必要があるからである．以下，Web ページなどを構成するマークアップ言語のデータは 11%を占めており，論文などの主要なフォーマットである PDF は約 9%となる．

図 4-5 は DJ に登録されたデータの種類である．数値データが最も多く，663 件の DJ のデータは数値を含んでいるデータである．また，583 件の DJ はテキストデータを含んでいることも分かる．データの種類は保存形式と密接な関係があり，CSV で保存されているデータは数値，テキストが表形式で保存されている．また，統計データの多くは日付の情報を含んでいることが多いため，時系列の形を取っている．その他，画像やグラフなどのデータは，報告書などに含まれているものを表している．

図 4-6 は DJ に登録されたデータの共有条件の内訳を表す．図にあるように，行政や公的機関が公開するオープンデータなどの共有可能データの登録数が最も多く，495 件のデータが DJ として登録されている．一方で，商用データや共有には交渉などの条件が必要となる一般に共有できないデータの DJ の登録数は共有可能データと比較するとやや少ない. また，データの共有条件が記載されていないものは 71 件ある[21]．例えば，「研究目的においてのみ販売する」など，一つのデータで複数の共有条件を持っているも存在するため，重複を含んでいる．条件付き（交渉等が必要）なデータであるデータが最も多く，123 件存在した．続いて範囲を限定して共有可能なデータ 86 件となった．71 件は社内での利用を想定しているなどの理由により共有不可能なデータとなっている．14%は，共有できないが詳細な条件については未定であるデータとなっている．また，注目すべきなのは，すでにデータ市場において価格が決定しているデータは 52 件（全体の 13%）あり，売買により一定の価値が認められているものも存在することが分かる．新聞記事のデータ[22]やテレビ番組のメタデータ[23]などがすでに販売されている．

[21] データジャケットの記述は，データ保有者が記入可能な情報のみを記入する形式となっているため，いくつかのデータジャケットには共有条件が記載されていないことに留意いただきたい．

[22] 日本データベース開発株式会社（http://www.ndk.co.jp/）など

[23] 株式会社エム・データ（http://mdata.tv/）など

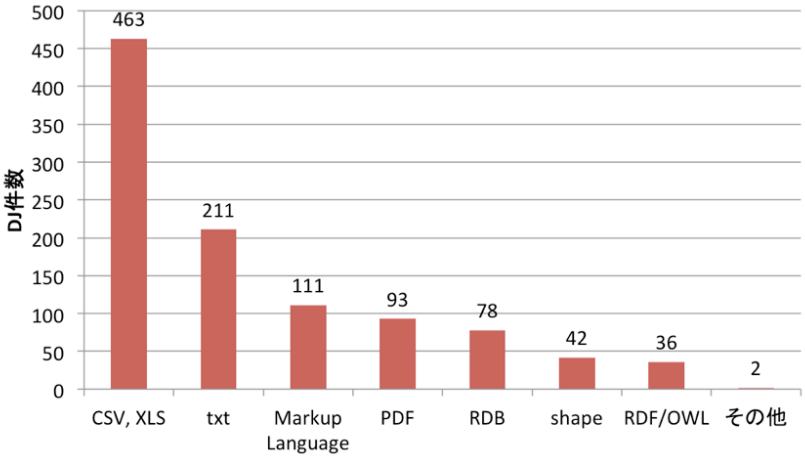

図 4-4　DJ に登録されたデータの保存形式

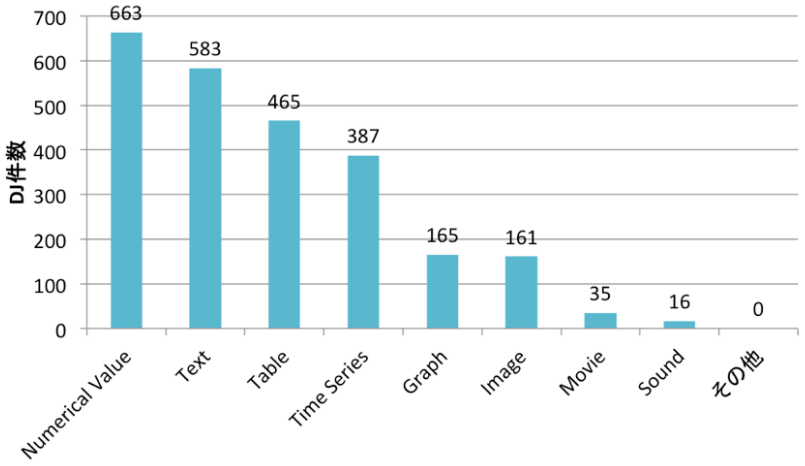

図 4-5　DJ に登録されたデータの種類

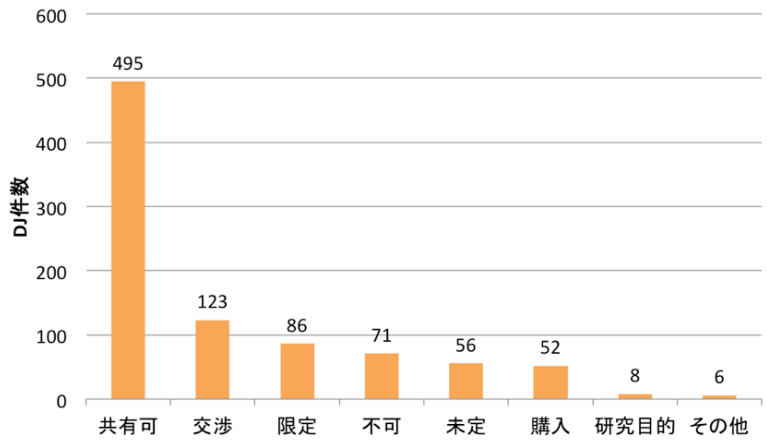

図 4-6　DJ に登録されたデータの共有条件

4.3.3 RDF による構造的記述

4.1 節にて説明したように，DJ はデータに関する 12 項目の情報（データ概要，変数ラベル，共有条件，分析プロセスなど）を有している．データが保有する変数ラベルやその件数はデータによって様々であるため，表形式のデータ記述では，他の DJ と列名を揃える必要があり，1 つのセル内に複数の情報を記述しなければならないこともあり得る．そのため，RDB や CSV などの表形式のデータ構造化は適当ではない．

また，DJ はデータ利活用方法検討ワークショップ IMDJ での利用を想定しており，DJ は組み合わされた「ソリューション」，そしてそのソリューションが満たした「要求」との連結を考慮したデータ記述方法である必要がある．なお，IMDJ によるデータ利活用案検討後は，シナリオマップ上に図 4-7 のように付箋が貼られた状態となる．このようなレイヤーの異なる複数の情報をデータベースに格納することは容易ではない．表形式のデータ記述では，要求，ソリューション，DJ の 3 つのデータベースを構築した後に，データベースを横断したリソース同士の関係を表現し，的確なデータを参照できる必要がある．このように，種類の異なるデータ同士のつながりやデータの組み合わせによって創出されたソリューション及び要求のようにネットワーク構造を取り得るリソースを構造化するためには，CSV や RDB などによる表形式の記述ではなく，RDF によるグラフ形式をベースにした記述が適切であると考えることができる．

図 4-7　IMDJ 後のシナリオマップの様子（黄色付箋は消費者の要求，青色付箋はソリューション，赤色付箋は追加の DJ を表す）

RDF（Resource Description Framework）はセマンティック Web の基盤となる言語の枠組みであり，ラベルと方向性のあるグラフベースの記述言語である．RDF は主語，述語，目的語を基本的な記述単位とし，この組み合わせによりネットワーク状の知識や概念，リソースの関係を表現できる（Berners-Lee, 2001; Yu, 2011）．RDF のグラフ形式の記述は，項目が増えてもノードとリンクを追加することでデータに含まれる情報を追加可能であり，情報の追加により構造の変更を余儀なくされる可能性のある表形式のデータベースの欠点を補うことができる．さらに，URI（Uniform Resource Identifier）によりデータのリソース名が一意に決まるため，参照するリソースが自然言語による意味のゆらぎに左右されることはないため，データベースが複数存在しても，参照すべきデータを特定することが可能となる. 以上の特徴により, RDF による記述は計算機による可読性を高めることが可能となる．

RDF の記述規則に基づいて，表 4-1 に示したように各 DJ 記述項目に対応した述語を設定した．なお，RDF で用いる語彙は既存のものほど再利用性が高まるため，セマンティック Web の文脈では Dublin Core（メタデータ記述の核となることが認められた語彙）のように意味が共有された語彙を用いることが望ましいとされる．しかし，本研究では次節で実装するデータ利活用知識の論理的な関係を実装するフレームワークとして RDF を採用しているため，一部の述語に独自の語彙を設定した．

表 4-1　DJ 記述項目と対応する RDF の述語

記述項目	述語
DJ の ID（自動で付与）	dj:id
データのタイトル	rdfs:label
データ概要	dj:outline
データの所有者とその所在	dj:ownership
データ収集方法やコスト	dj:collecting_cost
データの共有条件	dj:sharing_policy
データの種類	dj:type
データの保存形式	dj:format
変数ラベル	dj:variable
分析・シミュレーションプロセス	dj:analysis
分析・シミュレーションプロセスの結果	dj:outcome
分析・シミュレーションプロセスの結果以外に期待する分析	dj:anticipation
自由記述（データに関する補足事項）	dj:comments

図 4-8 は表 4-1 で設定した述語を用いて RDF 記述をグラフ表現したものである．ノードだけでなくエッジにもラベルが付いており，それぞれ意味を有している．グラフの中心の楕円のノード dj:0801 はグラフの主語，すなわち DJ の本体を表す．そしてエッジの述語は主語のリンク先の目的語との関係を表現している．例えば，dj:0801 から伸びているエッジ rdfs:label は，目的語が「データのタイトル」であることを表現する述語である．つまり，主語である dj:0801 のタイトルは「購買履歴データ」であることを意味している．また，述語 dj:variable で表されるのは DJ の変数ラベルであるが，dj:0801 は「氏名」，「性別」，「顧客 ID」，「支払金額」，「購入品目」の変数ラベルを有していることが理解できる．また，RDF のグラフ形式の記述では，エッジとリンクを追加することでデータに含まれる情報を追加可能であるため，新たに公開可能な変数ラベルがあとから追加されたとしても，述語と目的語を記述するだけ完了し，新たなデータの追加によりデータ構造を変更する必要がない．なお，本研究では RDF の記述を XML 形式（RDF/XML）で記述し，sparqlEPCU[24]を RDF ストアとして用いる．また，図 4-8 のグラフを RDF/XML で記述した例を図 4-9 に示す．このように，RDF では主語，述語，目的語の組み合わせをトリプルと呼び，基本的な表現単位としている．これらを組み合わせることにより，ネットワーク状の知識や概念，リソースの関係を柔軟に表現できる．

次節では，本設で検討した DJ の RDF を用いた構造的記述を用いたデータ利活用知識の構造化について議論する．IMDJ にて創出された要求及びデータ利活用案（ソリューション）をモデル化し，RDF を用いてデータ利活用知識として知識ベースを作る．そして，構造化したデータ利活用知識の再利用し，データについての情報を検索する DJ ストアというアプリケーションを実装し，その性能を実験的に評価する．

[24] http://lodcu.cs.chubu.ac.jp/SparqlEPCU/

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix dj: <http://datajacket.org/datajacket/>.
```

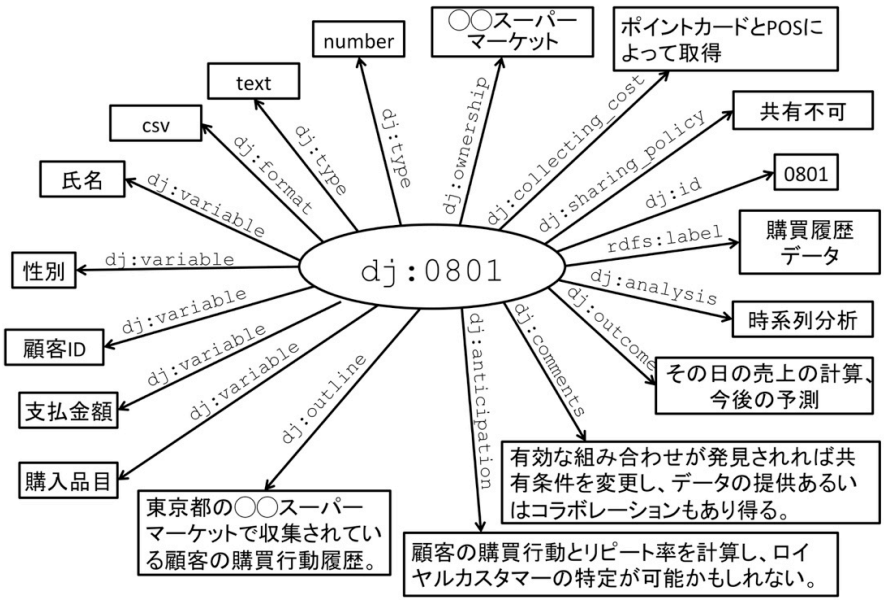

図 4-8　DJ の RDF のグラフ形式での表現例

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:dj="http://datajacket.org/datajacket/"
>

<rdf:Description rdf:about="http://datajacket.org/datajacket/dj0801">
        <rdfs:label>購買履歴データ</rdfs:label>
        <dj:id>0801</dj:id>
        <dj:outline>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。</dj:outline>
        <dj:ownership>◯◯スーパーマーケット</dj:ownership>
        <dj:collecting_cost>ポイントカードとPOSによって取得</dj:collecting_cost>
        <dj:sharing_policy>共有不可</dj:sharing_policy>
        <dj:type>number</dj:type>
        <dj:type>text</dj:type>
        <dj:format>CSV</dj:format>
        <dj:variable>氏名</dj:variable>
        <dj:variable>顧客ID</dj:variable>
        <dj:variable>支払金額</dj:variable>
        <dj:variable>購入品目</dj:variable>
        <dj:analysis>時系列分析</dj:analysis>
        <dj:outcome>その日の売上の計算、今後の予測</dj:outcome>
        <dj:anticipation>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。</dj:anticipation>
        <dj:comments>有効な組み合わせが発見されれば共有条件を変更し、データの提供あるいはコラボレーションもあり得る。</dj:comments>
</rdf:Description>

</rdf:RDF>
```

図 4-9　DJ の RDF の XML 形式の記述例

4.4 データ利活用知識の構造化と検索システム

本節では，データの有用性検討のためのデータベースとデータの構造化について検討し，ユーザーのデータ利活用方法検討を支援する検索システム Data Jacket Store（以下，DJ ストア）を実装，その性能を評価する．データジャケットに含まれる情報だけでなく，過去に検討された要求，データ利活用案によって有用性を認められたデータの関係のモデルを構築する．このモデルは知識構造化において非常に一般的な関係であるにもかかわらず，データ利活用に関する従来研究では十分に検討されてこなかった．提案モデルに基づき，データ利活用知識を構造化し，データ利活用方法検討支援の可能性について実験的に評価したことが本論文の新規性である．本システムにより，データジャケットを利用する人が自分と異なる視点を持つ他のユーザーが考案した該当データの使い道を発見したり，逆に，ある用途を検索しようとする人が過去の別の人が考案したデータ結合案に注目することによって役に立つデータについての情報を探し出すことが可能となる．

4.4.1 Data Jacket Store の設計

第 2 章及び前節にて説明したように，企業や組織のデータ利活用への期待は高く，現在登録されているデータジャケットは増加傾向にあり，今後ますます増えることが予想される．すると，データジャケットは公開されているものの，どのデータジャケットが自分と関係があるのか特定できないという問題が発生し得る．つまり，情報過多によりデータに関する情報が参照できる状態にあるにも関わらず，必要な情報を入手することが困難な状況が生じる可能性がある．人間が多くの情報から的確に必要な情報を引き出し，問題解決に必要な知識の組み合わせを発見することが困難であることは，人間の認知の限界の観点から指摘されており（Simon, 1955），膨大な情報からユーザーの欲している情報を検索し，推薦するシステムは今まで多くの研究や提案がなされてきた（Goldberg et al., 1992; Herlocker et al., 2004; 神嶌, 2006 など）．そこで，データジャケット関連技術として，ユーザーにとって有益な情報を見つけ出す検索・推薦システムが必要と考えることができる．膨大なデータについての情報（データジャケット）から的確に必要な情報を引き出す検索・推薦システムが実現すれば，データに基づく意思決定や効率の飛躍的向上が期待できる．

しかし，世の中にどのようなデータが存在し，どこにあるのかを知る方法は意外に少ない．Web 上に存在するデータであれば自然言語による検索が可能かもしれないが，欲しいデータを的確に検索し入手することは困難である．この問題の原因として，人間にとって自分の関心事をデータ内の用語（変数名や概要など）で表すことは必ずしも容易ではないため，計算機にとって人の関心事を理解することが難しいことが挙げられる．文字やデータをただ羅列するのではなく，計算機に可読な記述と，人の関心を表す記述を結ぶように

知識やデータの関係を構造化すれば，知識として再利用が可能となる．

人が関心を表す形態のうち，データのみならず利用者が問題として認識している「要求」とその解決方法「ソリューション」の再利用化が有用であると考えられる．欲しいデータが具体的に決まっているユーザーには，検索文章と一致する情報（データジャケット）の検索が有効に作用するかもしれない．しかし，欲しいデータが具体的に決まっていないユーザーの場合，様々な情報を閲覧し，比較検討する中でデータに関する知識を獲得し，必要な情報を選択すると考えられる．後者の場合は，自分にとって価値ある情報を得るまでに，漠然とした自分の要求やアイデアを述べ，過去に考案されたデータ利活用案を照合すると考えられる．すなわち，データのみならず多様な文脈とその背景にある様々な人の観点から示されたデータ利活用案である「ソリューション」及び「要求」を構造化し，再利用することで，自分の要求に合うデータがどこにあるのか，そして，自分のソリューションがどのようなデータ，分析方法を用いることで実現し得るのかについて知ることができる．本節の研究では，以上の機能を実装したシステムを DJ ストアと名付け，データ利活用知識構造化と検索システムの設計及び実装方法について議論する．

データ利活用知識として再利用し，追って構造化される「要求」及び「ソリューション」はデータ利活用方法検討ワークショップ IMDJ において創出されたものを用いる．

4.4.2 データ利活用知識のモデル

IMDJ で創出された「要求」，「ソリューション」及び「データ」の関係をデータ利活用知識として再利用可能にするために，知識を記述するモデルについて検討する．実験的データ市場において保有者は「データ」，利用者は「要求」，提案者は「ソリューション」を提供すると考えると，データ利活用知識の最小単位は以下の 2 つが考えらえる．

- データ利活用知識 1：あるソリューション（$solution$）はあるデータ（$data$）を用いている
- データ利活用知識 2：あるソリューション（$solution$）はある要求（$requirement$）を満たす

この 2 つの知識は，主語，述語，目的語で表現可能な構造を成しているので，「あるソリューションはあるデータを用いている」ことを表す述語（$\boldsymbol{combine}$）及び，「あるソリューションはある要求を満たす」ことを表す述語（$\boldsymbol{satisfy}$）を定義し，モデル化できる（図 4-10）．本論文のデータ利活用知識は IMDJ で創出された要求，ソリューション，データジャケットの連結によって表すため，データ利活用知識 1 のデータは，以下データジャケットを指すこととする．計算機の可読性向上と実装を考慮し，これらを二項の述語論理で記述すると，データ利活用知識 1 はモデル(4.4.2.1)，データ利活用知識 2 はモデル(4.4.2.2)と表せる．

$$\boldsymbol{combine}(solution, data) \quad (4.4.2.1)$$

$$\boldsymbol{satisfy}(solution, requirement) \quad (4.4.2.2)$$

これらのモデルを組み合わせることで，より複雑なデータ利活用知識を構造的に記述することが可能となる．例えば，2 つのデータ（$data1$及び$data2$）を用いて創出されるソリューション（$solution$）の場合は，$\boldsymbol{combine}(solution, data1 \wedge data2)$と記述できる．また，複数のソリューション（$solution1$及び$solution2$）が 1 つの要求を満たしている場合は，$\boldsymbol{satisfy}(solution1 \vee solution2, requirement)$と記述でき，複数のソリューションの組み合わせによって 1 つの要求が満たされている場合は，$\boldsymbol{satisfy}(solution1 \wedge solution2, requirement)$と表せる．つまり，上記のモデルを複数組み合わせることで，要求，ソリューション，データのデータベースを連結できる．例えば，要求のデータベースにクエリを発行すれば，該当する要求を満たすソリューション及びそのソリューションを構成するデータジャケットが発見可能となる．また，データジャケットのデータベースにクエリを発行し，そのデータジャケットを用いたソリューションについて知ることができる．過去のデータ利活用知識を構造化してシステムに実装することで，新たなデータ利活用検討の場にて過去に提起された要求に近い要求に対して，ある文脈からどのようなソリューションが提案でき，その文脈ではどのデータが利用可能かという情報が参照でき，個人では気づき得なかったデータの発見を促すことができると考えられる．

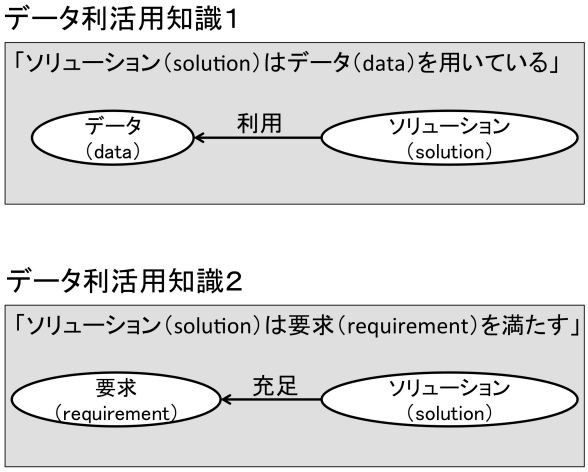

図 4-10　データ利活用知識モデルの図式化

4.4.3 検索クエリと DJ の取得

IMDJ におけるデータ利活用方法を検討する議論では，「安全な道路・交通システムを作りたい」といった要求や「街路灯の明るさを地図上に表示し，安全・安心なルートを検索

できるアプリケーションの開発」などのソリューション及びデータジャケットが自然言語にて創出される．こうして創出された「要求・ソリューション・データジャケット」という連結の内容を蓄積したデータベースが DJ ストアである．DJ ストア内では，まずユーザーが入力した文章から単語を抽出する．そして，これらの単語の OR 結合と，要求・ソリューション・データジャケットのデータベースを照合することで，関連するデータジャケット一覧を取得する仕組みとなる．つまり，ユーザーの入力文章をN個の単語からなるTとし，ある単語$t_i(1 \leq i \leq N)$を含むデータジャケットの集合を$DJ_{dj(t_i)}$とすると，取得されるデータジャケットの集合（DJ_{dj}）は式(4.4.3.1)となる．

$$DJ_{dj(T)} = DJ_{dj(t_1)} \cup DJ_{dj(t_2)} \cup \cdots \cup DJ_{dj(t_N)} = \bigcup_{i=1}^{N} DJ_{dj(t_i)} \qquad (4.4.3.1)$$

同様に，Tに適うソリューションに関連するデータジャケットを検索する場合，Tに含まれる各語 t_iを含むソリューションを取得し，T中のすべての語についてのソリューションと連結したデータジャケット一覧（DJ_{sol}）を返す．要求からデータジャケットを検索する場合は，t_iを含む要求を取得し，T中のすべての語について該当する要求を満たすソリューションと連結したデータジャケットの一覧（DJ_{req}）を返す．以上の仕組みにより，ユーザーは自身の入力文章Tに対して，関連するデータジャケットの集合（$DJ_{dj(T)} \cup DJ_{sol(T)} \cup DJ_{req(T)}$）を取得できる．データ利活用知識のモデルから構築した要求，ソリューション及びデータジャケットのデータベースを連結した概念図を示す（図 4-11）．また，図 4-12 は検索により取得されるデータジャケットの集合関係を表す．

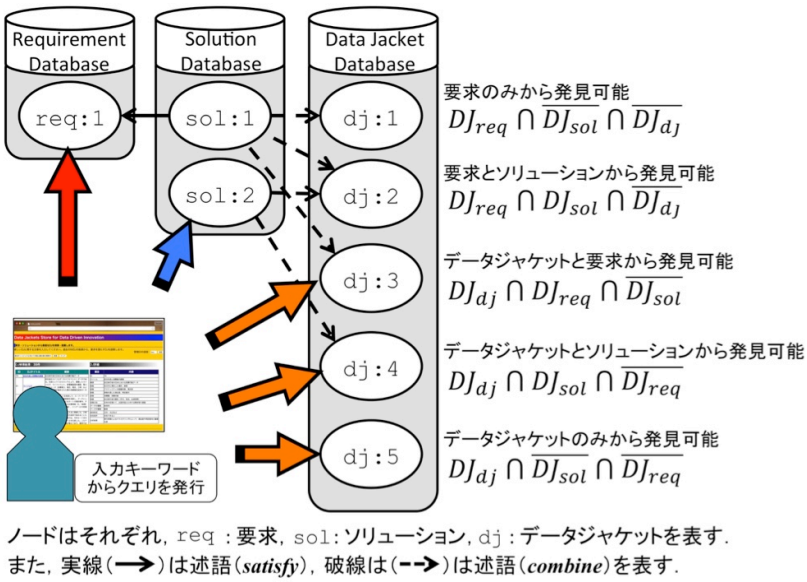

図 4-11　要求，ソリューション，データジャケットのデータベースを連結した概念図（早矢仕・大澤, 2016 より引用）

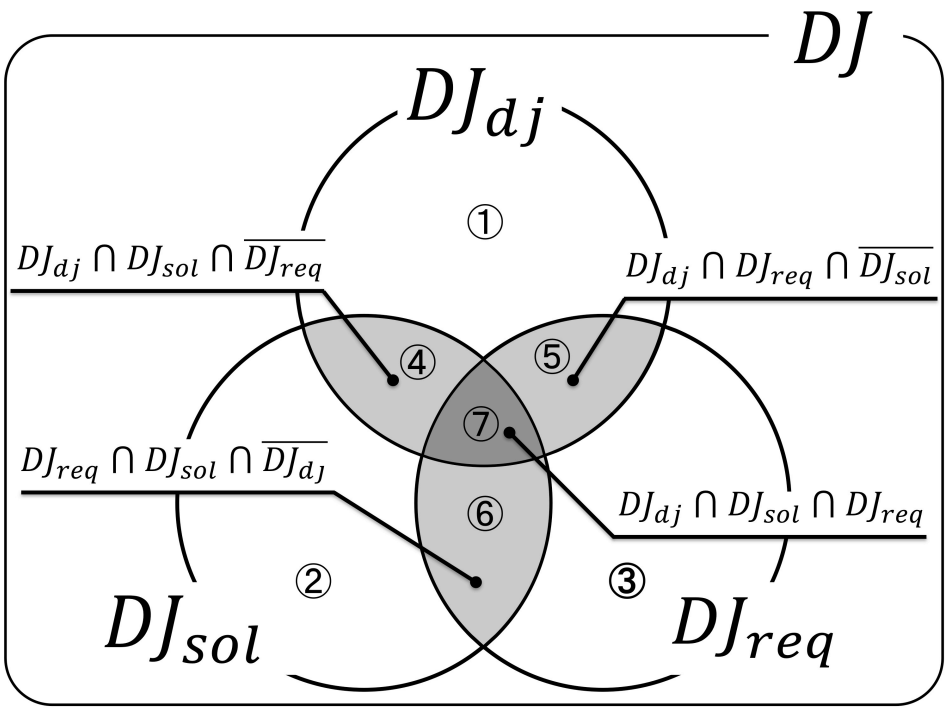

図 4-12　データジャケットの集合関係（$\boldsymbol{DJ_{dj}}$：あるキーワードを含むデータジャケットの集合，$\boldsymbol{DJ_{sol}}$：あるキーワードを含むソリューションに用いられたデータジャケットの集合，$\boldsymbol{DJ_{req}}$：あるキーワードを含む要求を満たしたソリューションを構成するデータジャケットの集合）

4.4.4 Data Jacket Store の実装

データ利活用知識を再利用し，ユーザーにとって有益な情報を引き出す検索システムの実現には，モデルを実装する適切なフレームワークの選択が重要である．モデル(4.4.2.1)及びモデル(4.4.2.2)で表されるデータ利活用知識は，主語，述語，目的語を持つグラフ構造を有しているため，4.3 節にて構築した RDF による記述を拡張して用いる．RDF の構文に従って，ソリューションとデータジャケットを繋ぐ述語（***combine***）を `sol:combine`，ソリューションと要求を繋ぐ述語（***satisfy***）を `sol:satisfy` と定義することで，4-10 に示した構造を RDF ストアとして実装できる．RDF で記述された情報は，クエリ言語 SPARQL を用いてグラフパターンを検索し，要求から，その要求を満たしたソリューションを構成するデータジャケットなどが取得可能となる．図 4-13 は図 4-10 で外観を示したデータ利活用知識モデルを RDF で表した例である．楕円形ノードは左から要求（`req:0011`），ソリューション（`sol:0292`），データジャケット（`dj:0079` 及び `dj:0171`）を表す．要求からは，要求の ID を表す述語 `req:id` と要求の内容を表す述語 `rdfs:label` によって，`ID` 及び内容の情報が繋がっている．要求のノードは `sol:satisfy` によってソリューションを表す楕円形ノードと連結している．これはデータ利活用知識 1（モデル(4.4.2.1)）を表して

いる．ソリューションのノードからも要求と同様に ID を表す述語 sol:id 及びソリューションの内容を表す述語 rdfs:label が繋がっている．ソリューションからは，組み合わせた 2 つのデータジャケットが述語 sol:combine によって連結している．これはデータ利活用知識 2（モデル(4.4.2.2)）の組み合わせである．楕円形ノード dj:0079 及び dj:0171 はタイトルを表す rdfs:label，ID を表す dj:id，概要説明を表す dj:outline によって，それぞれの要素が繋がれている．なお，本来であればデータの組み合わせを表す述語 sol:combine をモデル(4.4.2.1)及び(4.4.2.2)に基づいて AND や OR 結合を考慮し sol:and_combine や sol:or_combine などと分けて記述するべきであるが，本研究では検索における複雑さを回避するため，RDF のデータモデルによってデータ・ソリューション・要求の関係記述には，すべて OR 結合を用いた．

また，データジャケットは，タイトル，概要説明，データの収集方法，変数名，保存形式，共有条件など 12 の項目を定義し RDF で記述しているが，本節で扱うデータ利活用知識モデルでは要求及びソリューションの名前と ID，データジャケットのタイトル，ID 及び概要説明のみを利用しているため，その他の要素は図 4-13 では省略している．なお，本研究では RDF をデータ利活用知識の論理的な関係を実装するフレームワークとして採用しているため，一部の述語に独自の語彙を設定し，要求，ソリューション及びデータジャケットを記述している．

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix   dj: <http://datajacket.org/datajacket/>
@prefix  sol: <http://datajacket.org/solution/>
@prefix  req: <http://datajacket.org/requirement/>.
```

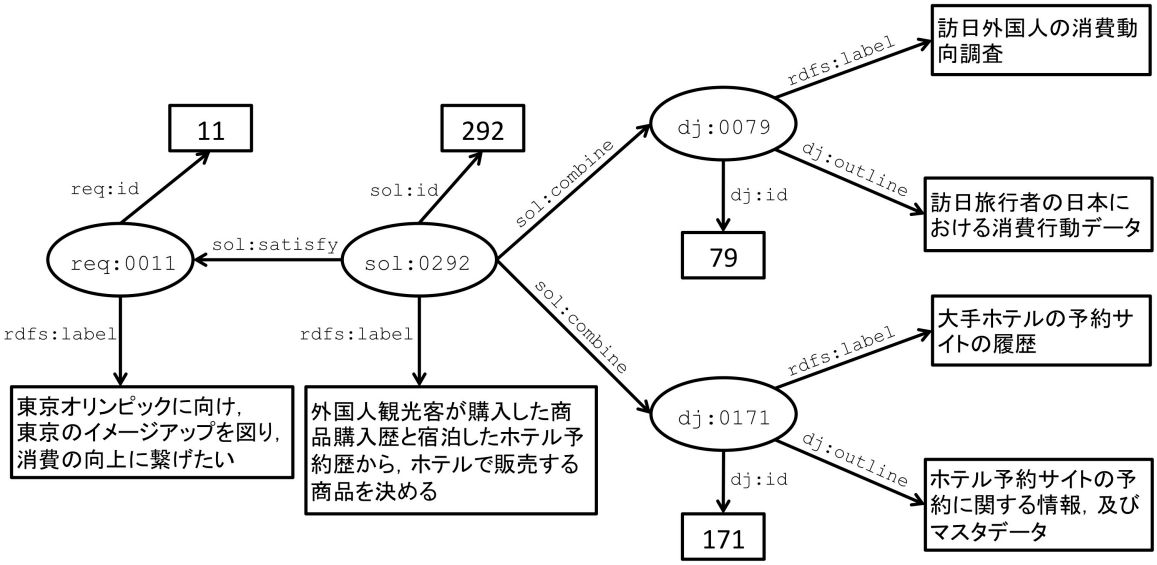

図 4-13　データ利活用知識モデルの RDF グラフによる表現例

図 4-14 は DJ ストアのユーザーインタフェースである．RDF ストアには，sparqlEPCU を用いた．さらに，入力文章を単語に分割する際には TinySegmenter[25]を用い，「データ」や「情報」などの極端に出現頻度の高い名詞，助詞，助動詞，記号類は不用語として除外した．例えば，「東京オリンピックに向けて外国人誘致のための新ビジネスを創出したい」という要求を入力すると，「東京」，「オリンピック」，「外国人」といった抽出単語から RDF ストアの `rdfs:label` 及び `dj:outline` で繋がれている情報を検索する．図 4-13 に示すように，「東京オリンピックに向け，東京のイメージアップを図り，消費の向上に繋げたい」という要求と，「外国人観光客が購入した商品購入歴と宿泊したホテル予約歴から，ホテルで販売する商品を決める」というソリューションが構造化されていると，そのソリューションに連結したデータジャケット「訪日外国人の消費動向調査」と「大手ホテル予約サイトの履歴」が得られる．検索文章内の単語とデータジャケット内の単語の一致だけではなく，過去に検討されたソリューション及び要求を介してデータジャケットを発見することができる．取得したデータジャケットは図 4-14 の画面左に検索結果として表示され，各データジャケットに記述された中身に相当する詳細情報はタイトルをクリックすることで，画面右に表示される．

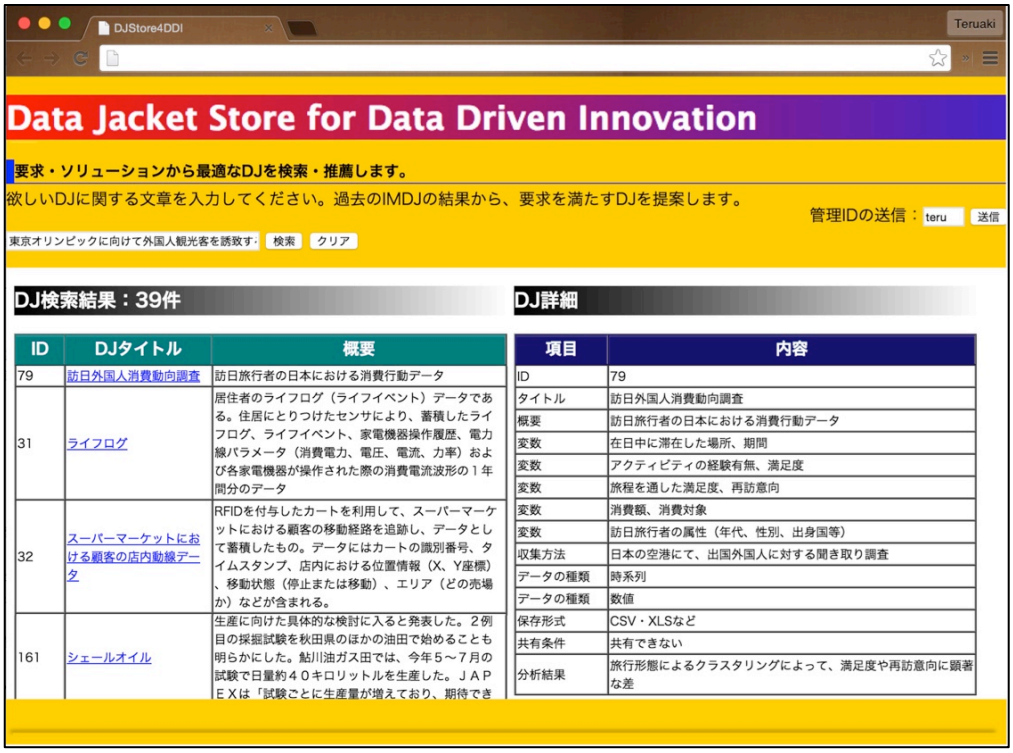

図 4-14　実験で用いた DJ のユーザーインタフェース（早矢仕・大澤, 2016 より引用）

[25] http://chasen.org/~taku/software/TinySegmenter/

4.4.5 Data Jacket Store の性能実験

データ利活用知識モデルを実装したシステムDJストアによって検索されたデータジャケットがデータ利活用方法の検討においてユーザーの期待を満たすかどうかを評価するための実験を行う．本システムの性能を評価する上で以下の2つの指針が重要である．

① ユーザーに提示されたデータジャケットがソリューション創出に役立ったか

② 提示されたデータジャケットが，ユーザーの潜在的な要求に適合するデータジャケットであるか

①については，検索されたデータジャケットのIMDJ中の利用率（どの程度ソリューション創出に利用されたか）を調べることで評価可能である．一方，②を検証するために，検索ごとにユーザーにヒアリングすることは現実的ではない．なぜなら，実験者の介入により参加者間のデータ利活用についての議論やIMDJの進行が阻害されてしまう可能性があるからである．そこで本研究では，DJストアの検索結果と選択されたデータジャケットのログを利用し，システムを評価するため，以下の3通りの発見データジャケットの集合を設定する．

- データジャケットに含まれる単語から発見されるデータジャケット集合（DJ_{dj}）
- ソリューションに含まれる単語から発見されるデータジャケット集合（DJ_{sol}）
- 要求に含まれる単語から発見されるデータジャケット集合（DJ_{req}）

データジャケットを包含する集合と各集合に含まれるデータジャケット数の比較により，検索文章内の単語との直接一致によるデータジャケット集合ではなく，要求及びソリューションを介して検索されたデータジャケット集合の利用期待度の高さが検証可能と考えられる．本実験では，IMDJを5回実施した．各IMDJの参加者は学生及び社会人である（総参加者は48人）．本実験ではIMDJを実施する机にノートパソコンまたはタブレット上で操作可能なDJストアを設置し，参加者がIMDJの最中に関連するデータジャケットを検索できる環境を構築した．特に，DJストアによって発見されるデータジャケットに対するユーザーの利用期待度の高さを調べるために，ユーザーがDJストアによって発見したデータジャケットをIMDJのシナリオマップ上に追加し，ソリューション創出に用いることを許可した．なお，各IMDJは90分間実施した．IMDJの実施手順の概要は以下である．

1. 事前にIMDJに用いるデータジャケットを20から25件の範囲で準備し，データジャケット間で共有された単語を結ぶグラフからマップを作成した（参加者の共通参照情報としてKeyGraph（Ohsawa et al., 1998）を用いたが，組み合わせのヒントの提示が趣旨であり，全5回の実験で統一して用いたので，可視化ツールの効果比較は論じない）．
2. 参加者は実在のデータ利用者の利害を主張する立場から要求を提示し，提案者の立

場から，データジャケットを組み合わせることで要求を満たすソリューションを創出する．

3. 利用者は自身の要求を満たすソリューションが創出された場合，ゲーム中の架空通貨を提案者に支払う（ソリューションの評価として計上する）．
4. 手順 1 から 3 を時間内に繰り返し，最も架空通貨を得た者が提案者としての勝者となる．また，利用者は購入したソリューションによって自身の満たされた要求を発表する．最も有用と認められた発表をした者が利用者としての勝者となる．

本実験において DJ ストアに格納したデータジャケットは 376 件，過去の IMDJ 情報として 296 件の要求，264 件のソリューションを含む総トリプル数（RDF における主語・述語・目的語の三つ組み構造）9,727 の RDF/XML 形式のデータを利用した．また，ユーザーには DJ ストアが提示するデータジャケットが，図 4-12 のどの集合に含まれるものであるかは明示しないものとした．

4.4.6 結果と考察

全 5 回の IMDJ において，DJ ストアによる総検索回数は 84 回，累積表示データジャケット件数は 1,737 件であった．DJ ストアは，検索文章内の単語の OR 結合からデータジャケット・ソリューション・要求のデータベースを照合することで，関連するデータジャケット一覧を取得する．しかし，データベースに該当するデータジャケットが含まれない場合，値は返ってこない．1 回の検索によってデータベースから値が返ってくる単語の数を有効単語数とし，有効単語数をデータベースごとに比較したところ，データジャケット，ソリューション，要求の各データベースにおける有効単語数はそれぞれ 145，127，133 であり，大きな差は見られなかった．検索文章内の単語によってアクセスするデータベースに偏りはないことから，以降の考察では各データベースへのアクセス回数ではなく，発見されたデータジャケット件数を比較することで DJ ストアの性能評価を行うこととする．

また，ユーザーに提示されたデータジャケット件数を図 4-12 における集合によって比較するために，DJ 群と Req-Sol 群を定義し，各データジャケットを包含する集合を表現する．

- DJ 群 : データジャケットに含まれる単語の一致によって発見可能なデータジャケットの集合（図 4-12 においてDJ_{dj}で表される集合．図中の①，④，⑤，⑦を併せた領域を指す）
- Req-Sol 群 : データジャケットに含まれる単語の一致では発見できないが，「ソリューションに含まれる単語との一致によって検出され，そのソリューションを創出するのに用いられたデータジャケット」または「要求に含まれる単語との一致から発見され，その要求を満たしたソリューションに用いられたデータジャケット」の集

合（図 4-12 において$DJ_{req} \cup DJ_{sol} \cap \overline{DJ_{dj}}$で表される集合．図中の②，③，⑥を併せた領域を指す）

4.4.6.1 提示データジャケット件数比較

図4-15はワークショップ中のDJストアを用いた被験者の検索によって発見されたデータジャケットの数をDJ群とReq-Sol群で比較したものである．データジャケットに含まれる単語の一致によって発見可能なデータジャケットの集合を対象とした検索により，DJ群のデータジャケットは 867 件発見された．そして，データジャケットに含まれる単語の一致では発見できないが，ソリューション及び要求を介して発見したデータジャケットの集合であるReq-Sol群を検索範囲に含めることで，さらに870件の関連するデータジャケットがユーザーに提示された．つまり，ソリューション及び要求を介することで，データジャケットのみを検索対象とした場合では発見できない情報が発見可能となることが分かった．

しかし，提示されたデータジャケットの件数を単純に比較するだけではユーザーにとって問題解決を行うのに役に立ったかどうかを判断することはできない．ユーザーに提示されたデータジャケットの利用期待度を評価するためには，閲覧回数の比較や，実際に要求を満たすソリューション創出に利用が期待できるとして IMDJ に追加された数及びソリューション創出に用いられた回数を集合別に比較することが必要である．続いてユーザーが詳細情報を閲覧した回数，IMDJ に追加情報として追加されたデータジャケットの数及びソリューションへの利用回数を比較する．

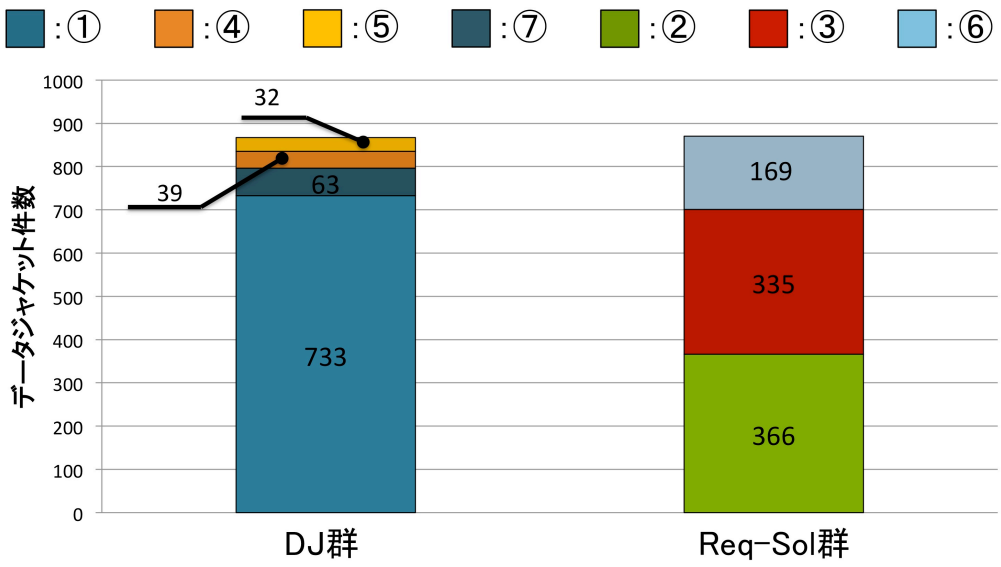

図 4-15　ユーザーに提示されたデータジャケット件数比較

4.4.6.2 ユーザーの詳細情報閲覧件数比較

DJ ストアにおいて検索が行われたのち，知識を獲得するためにユーザーがデータジャケットの詳細を閲覧したかどうかを調べ，包含される集合を比較した．これを調べることにより，ユーザーの検索により提示されたデータジャケットの中から，さらにどのようなデータジャケットに注目したのかを理解することができる．その結果，詳細情報が閲覧された DJ 件数は 67 件であり，そのうち DJ 群のデータジャケットは 38 件，Req-Sol 群は 29 件であった．つまり，DJ 群に含まれるデータジャケットの詳細情報閲覧が比較的多いことが分かった（図 4-16）．

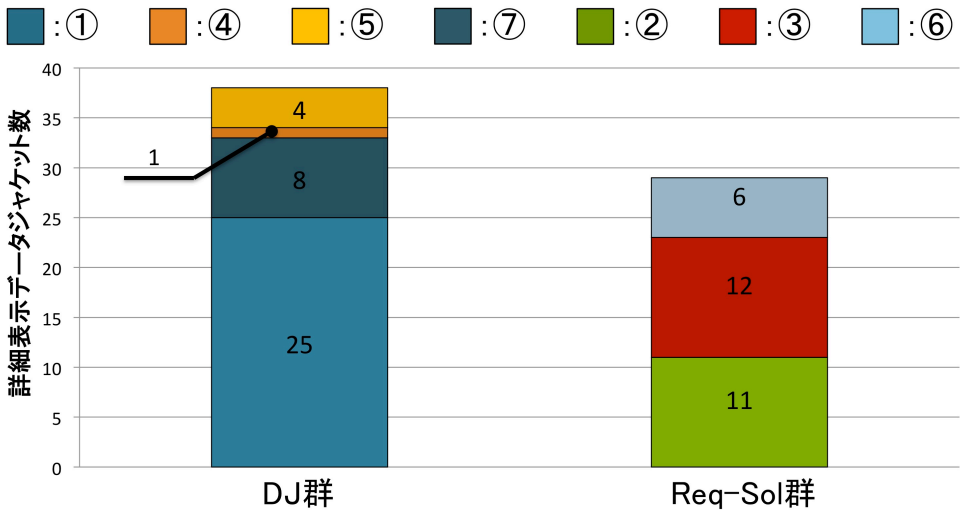

図 4-16　DJ ストアにおいてユーザーが DJ について詳細情報を閲覧したデータジャケット数の比較（図中の①から⑦は図 4-12 と対応している）

4.4.6.3 データジャケットついかとソリューション創出数，評価数の比較

各 IMDJ において追加され，ソリューション創出に用いられたデータジャケットを数えたところ，上述の DJ 群よりも Req-Sol 群の追加データジャケット数及びソリューションに用いられた件数の方が多いことが分かった（図 4-17）．さらに，データジャケットの利用回数及び IMDJ 中にソリューションが購入された回数（架空通貨による評価回数）も Req-Sol 群のデータジャケットを用いたソリューションの方が多いという結果を得た（図 4-18）．なお，図 4-17 及び図 4-18 の結果は，同じデータジャケットが異なるソリューション創出に用いられることによる重複を含む．ユーザーに提示されたデータジャケットの件数はほぼ同数であったにもかかわらず，必要な情報として追加され，ソリューション創出に用いられたデータジャケット数は Req-Sol 群の方が多いことは注目に値する．さらに，ソリューション創出に用いられたデータジャケットのうち，要求のみから発見可能なデータジャケット（図

4-12 の③の領域）は他と比較し複数回利用されていることが分かった．また，詳細情報が閲覧されたデータジャケットの件数は DJ 群の方が多いにもかかわらず，必要な情報として追加されたデータジャケット数は Req-Sol 群に含まれるデータジャケットであることも注目すべき結果である．

以上の結果から，要求及びソリューションを介して得られたデータジャケット（Req-Sol 群）はユーザーにとって高い利用期待度を持つという示唆が得られる．これは，データ内の用語をキーワードとした直接の一致によるデータジャケットではなく，過去に検討された要求とその解決方法（ソリューション）を介して得られたデータジャケット（Req-Sol 群）の方が，ユーザーが潜在的に欲している情報であり，ソリューション創出に有用である可能性を支持すると考えられる．また，要求のみから発見可能なデータジャケットは他の集合と比較して利用回数が多く，Req-Sol 群の中でも特に，要求から発見されたデータジャケットはユーザーの検索意図に合致し，利用期待度が高いことが示唆される．

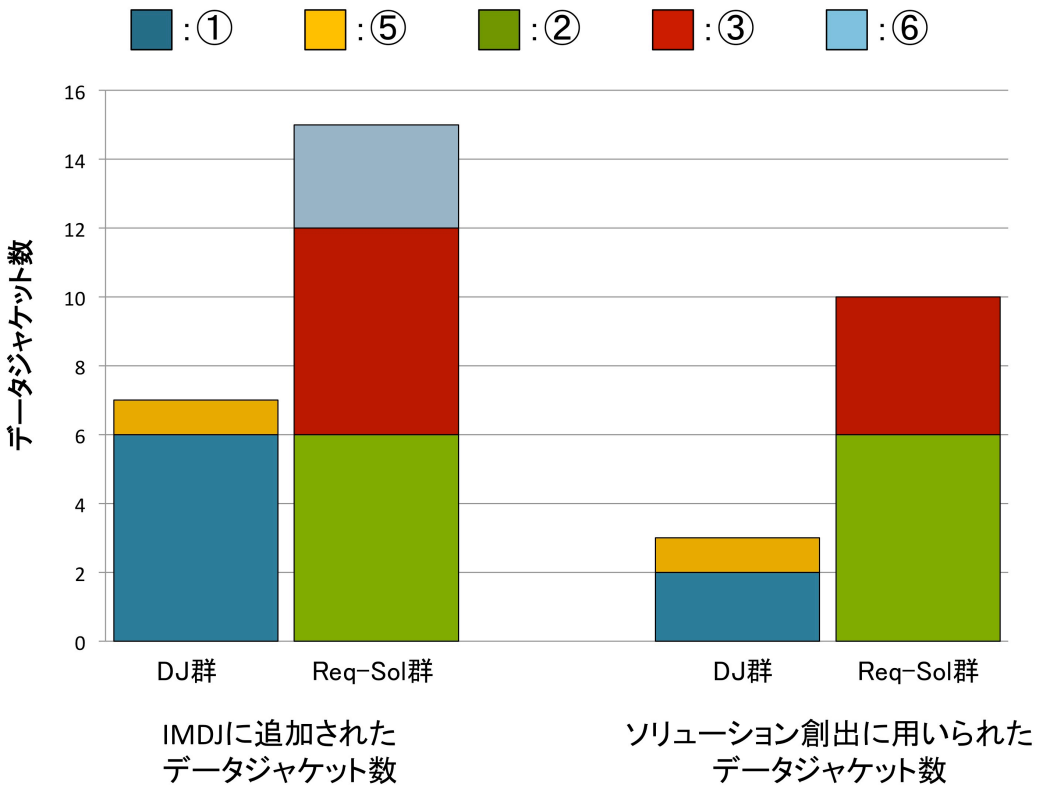

図 4-17　ユーザーが IMDJ において追加したデータジャケット数及び，ソリューション創出に用いたデータジャケット数

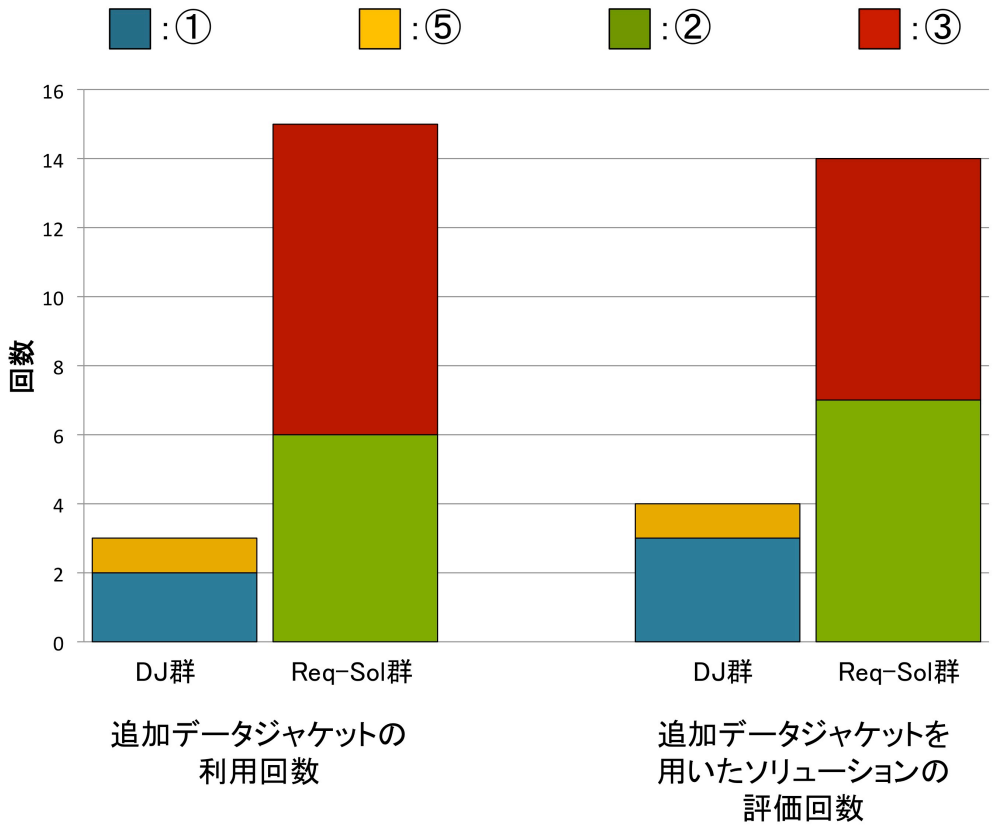

図 4-18　ユーザーがソリューションを創出するために利用したデータジャケットの利用回数及び，評価数の比較

4.4.6.4 既存データジャケットとの利用率と高評価率の比較

以上の議論では，検索によって追加されたデータジャケット集合を比較した．続いて，IMDJ のシナリオマップ上の既存データジャケットとの比較によって，追加されたデータジャケットの利用期待度を評価する．

IMDJ のシナリオマップには，予めテーマに基づいて収集されたデータジャケットが配置されている．全 5 回の IMDJ の既存データジャケットは 112 件（これを既存 DJ 群とする）であり，それらのデータジャケットのみを用いて創出されたソリューションは 88 件，評価されたソリューションは 60 件であった．一方，図 4-18 に示したように，追加データジャケットを用いて創出されたソリューションは Req-Sol 群において 15 件，DJ 群において 3 件である．また，評価されたソリューションは Req-Sol 群において 14 件，DJ 群において 4 件であった．評価されたソリューションとは，要求を満たしたとして架空通貨が支払われたソリューションの数を意味する．これらのソリューション数と DJ ストアの検索によって追加されたデータジャケットを用いたソリューション数を比較した．比較の指標として，データジャケット利用率（式(4.4.6.4.1)）とソリューション高評価率（式(4.4.6.4.2)）を導入した．利用率はデータジャケットがソリューションに利用された割合を指し，高評価率は創出さ

れたソリューションが評価された割合を意味する．式(4.4.6.4.1)及び(4.4.6.4.2)におけるソリューション創出数とは，ソリューション創出に用いられたデータジャケット数であり，評価ソリューション数とは，利用期待度が高いとして架空通貨が支払われたソリューションの件数である．なお，利用率及び高評価率のデータジャケット数には，同じデータジャケットが異なるソリューション創出に用いられることによる重複を含む．

$$\text{利用率} = \frac{1}{n}\sum_{k=1}^{n}\frac{(\text{ソリューション創出数})_k}{(\text{データジャケット数})_k} \qquad (4.4.6.4.1)$$

$$\text{高評価率} = \frac{1}{n}\sum_{k=1}^{n}\frac{(\text{評価ソリューション数})_k}{(\text{ソリューション数})_k} \qquad (4.4.6.4.2)$$

※ nは IMDJ の回数，kは IMDJ の添字を表す．

各群において利用率を比較すると，Req-Sol 群によるデータジャケットは既存 DJ 群及び DJ 群と比較し，IMDJ 中の利用率が高いことが分かる（図 4-19）．一方，高評価率においては DJ 群のデータジャケットを用いたソリューションが比較的高いという結果が得られた．しかし，DJ 群に含まれるデータジャケットの利用率は最も低く，ソリューションの評価は高くなる傾向はあるものの，安定的な利用には適さないと言える．一方で，Req-Sol 群に含まれるデータジャケットは利用率，高評価率ともに既存 DJ 群よりも高く，高評価を得るソリューションを創出できる可能性が高く，安定した利用が可能であると考えられる．

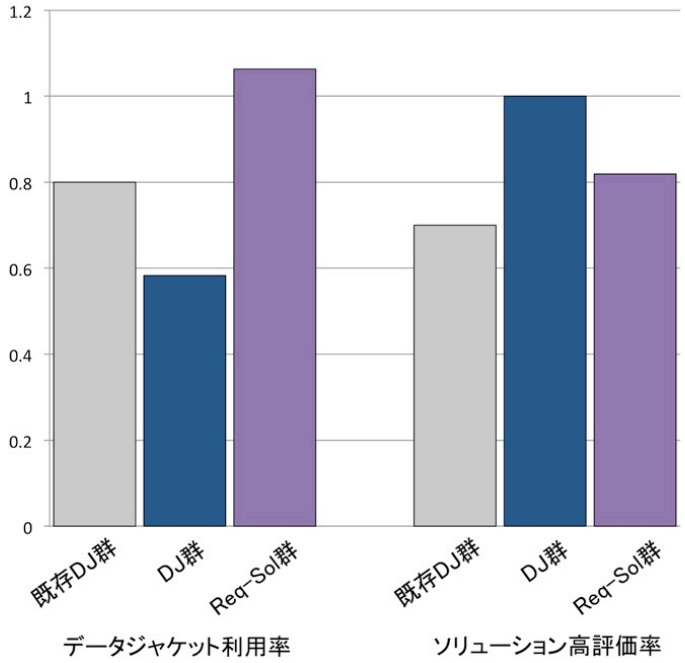

図 4-19　既存 DJ 群，DJ 群，Req-Sol 群の利用率と高評価率

4.4.6.5 データ利活用方法検討のプロセスと検索行動

続いて，DJ ストアによるデータジャケットの検索行動が行われた時間を調べた．約 90 分の IMDJ において，検索クエリを発行した回数及びデータジャケットのタイトルをクリックし，データの詳細を確認した回数を検索回数として図 4-20 に表す．以上を調べることにより，データ利活用方法検討プロセスにおいて，IMDJ 参加者がどの時点で外部知識が必要となるのか知ることができる．データ利活用方法検討プロセスを調べたところ，特に IMDJ 後半に DJ ストアを用いた検索が頻繁に行われていることが分かった．さらに，DJ 群及び Req-Sol 群で追加され，ソリューション創出に用いられた DJ の全てが，IMDJ の後半に検索され，追加されたものである．つまり，データ利活用方法検討初期は，シナリオマップ上のデータジャケットなどの既存の知識や情報でソリューションを創出するなどの対応が可能であるが，時間経過によって既存の知識や情報のみでは解に至れず，外部の知識や情報を求めて検索行動が頻繁に行われるようになったと考えることができる．以上から，データ利活用知識モデルを用いた知識の蓄積は DJ ストアの情報ソースとして必要不可欠であるが，新しく実施する IMDJ では既存データジャケットに縛られず，自由に要求や着想を表現しながら DJ ストアから検索されるデータジャケットを援用し，提案を進めていくという方法がデータ利活用方法の検討を効果的に進めるものと考えられる．

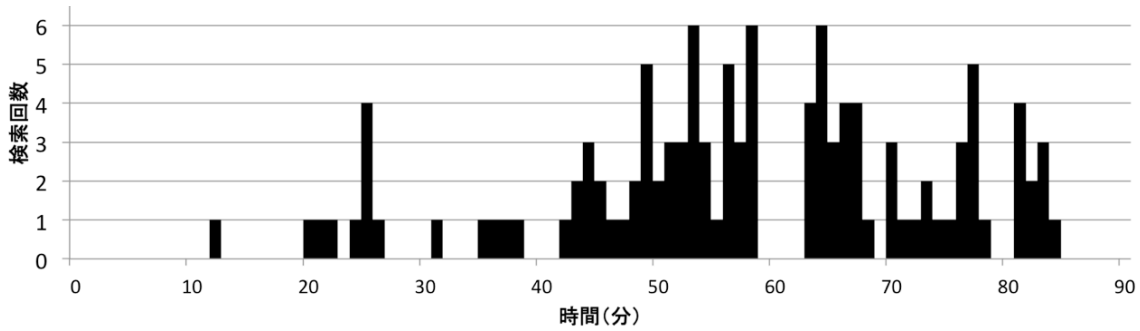

図 4-20　IMDJ における検索行動の時間と回数

4.4.6.6 データ利活用知識モデルの有用性についての考察

データ利活用知識 1（あるソリューションはあるデータを用いている）及びデータ利活用知識 2（あるソリューションはある要求を満たす）のモデルによって構築した複数のデータベースを検索対象に含めることで，問題解決においてユーザーの利用期待度の高いデータジャケットが検索可能であることが分かった．本実験では，要求やソリューションの情報を被験者には開示しないにもかかわらず，要求とソリューションを介した検索結果である Req-Sol 群のデータジャケット追加回数及び利用回数が DJ 群と比較して高いという注目す

べき結果を得た．ユーザーは問題解決において欲しいデータが具体的に決まっておらず，要求を満たし得るソリューション及びデータジャケットの繋がりについて知らなかったことが要因であると考えられる．なぜなら，Req-Sol 群のデータジャケットを発見したユーザーの検索文章を見ると，具体的なデータを指す文章ではなく，「健康長寿の延長」，「病気を治したい」，「Wi-Fi を気軽に利用できる場所がほしい」のように抽象度の高い文章が入力されていた．すなわち，確信を持って該当するデータについて詳細に表す文章ではなく，要求やソリューションに近い単語を含むクエリによって，ユーザーは自身が欲するデータジャケットを検索していたことが分かる．以上の考察から，要求及びソリューションのデータベースを検索対象に含めることで，検索行動時に欲しいデータが具体的に決まっていないユーザーの検索意図を満たすデータジャケットの発見が促されたと考えられる．

以上の考察は，アンケートによっても裏付けられる．本実験では IMDJ 終了後に参加者に自由記述のアンケートを実施した．アンケートの結果，「新しいデータの発見がソリューション創出に役立った」，「『安心』などの抽象度の高いワードによって面白いデータジャケットが抽出された」，「思いがけないデータの繋がりが見つかった」などの提示されたデータジャケットから新しい知識を獲得したという回答が見られた．つまり，これらのユーザーは具体的に欲しいデータを想定したり，確信のある文章の形で要求を表現して検索を行ったのではなく，抽象的な文章によるクエリを用いた検索によって様々なデータジャケットを比較検討する中でデータに関する知識を獲得したものと解釈できる．この結果より，データ利活用知識モデルによる要求及びソリューションを介した検索方法は，検索時に欲しいデータが具体的に決まっていないユーザーの検索意図に合致した情報の検索に有用であることが分かる．一方，DJ 群のデータジャケットを発見したユーザーは，「自動車の販売台数」や「企業間取引情報」という検索文章からも分かるように，欲しているデータに関する具体的なキーワードを入力したことが確認された．

続いて，検索によって得られたデータジャケットの利用方法が DJ ストアに格納された過去のデータ利活用知識と類似しているか否かについて調べた．例えば，「大手ホテル予約サイトの履歴データ」は「東京オリンピック時の高額短期雇用の創出」という労働者を対象としたソリューションに利用されていた．本実験でも東京オリンピックに関する検索文章から発見されたが，「旅行客の国別の交流の場の提供」という旅行者を対象としたソリューションに用いられた．過去に検討された利用方法とは異なる方法で利用されているが，データが利用された文脈及び要求については比較的類似している．一方，「ライフログ」などのデータは，過去に「高齢者の外出及び運動の促進」という文脈でソリューションが考案された．本実験による検索でも高齢者の健康とスポーツに関する検索文章から発見され，「高齢者の健康状態を考慮した外出先の推薦」という過去に検討された利活用案と類似し

たソリューションに用いられた．また，高齢者のニーズに対するデータ利活用という点から，要求も比較的類似していると言える．以上の例から分かるように，ユーザーは必ずしも過去のデータ利活用案と同じ方法でデータを利用しているわけではないが，過去に提起された要求と類似の要求を満たすソリューション創出に該当データを用いていることが分かった．以上の結果から，発見されたデータは多様な文脈で利用されるものの，ユーザーの検索意図を満たすデータジャケットの発見を促す上でデータ利活用知識の構造化は有効に作用したと考えられる．

4.4.6.7 まとめ

データの再利用環境の整備に注目が集まっている一方で，データをどう使うか，データを使うことによりどのような問題が解決可能であるのか，というデータ利活用知識を再利用する仕組みの検討は十分に行われてこなかった．組織横断的な協働を促進する知識構造化（白松ら, 2016）が考案されるなど，構造化したデータ利活用知識を再利用する仕組みが提案され始めているが，白松らの研究では，社会課題と解決目標及び関連する記事・データという社会課題を有する狭い領域及びコミュニティにおける知識を対象としている．一方，DJ ストアの知識構造化は，社会的課題に限らず社会に存在するあらゆるステークホルダーの多種多様な要求に対する解決策及び関連するデータを対象としている点で異なる．課題と解決策の知識構造化という共通の解決課題に取り組んでいることから，本研究のデータモデルを援用する方法も考えられるだろう．

一方，本研究の目的は，データ利活用知識のモデル化の有用性検証であったため，単語の重み付けによるランキングなどは扱わなかった．ソリューション創出回数やコーパスを用いた単語の重要度を考慮したシステムの拡張が今後の課題である．また，データ利活用はデータの組み合わせだけでなく，適切な分析手法との組み合わせによって実現する．Hayashi & Ohsawa（2016a）の研究では，分析ツールの概要情報（ツールジャケット）を用い，DJ に含まれる変数名の情報（変数ラベル）から適切なデータと分析ツールの組み合わせを議論可能とする知識構造化について扱っており，本節にて議論した DJ ストアの分析ツールの概要情報への拡張も可能であると考えられる．

4.5 共有条件に着目したデータの利用期待度

前節では，過去に検討されたデータ利活用知識の構造化と再利用は新たな知識獲得に有用であることが分かった．特に，実験的データ市場において既存の知識や情報だけでは解決できない問題に直面した際に，データに関する情報の陳列だけではなく，データ利活用知識の構造化による検索システムが利用価値の高いデータの発見と問題解決を促し，データに対する新たな需要を喚起する可能性が示唆された．

データ市場はオープンデータに代表される公開可能データだけでなく，個人や企業の秘匿データも利用方法を検討し，価値を見出されるイノベーションの場であることはすでに述べた．しかし，データの共有可能性を意味する共有条件はデータ市場においてどのように作用し得るのかは明らかとなっていない．本節では，データ市場における共有条件に着目し，どのようなデータの利用価値が高くなり得るのか，という問について，データの利用期待度に着目して考察を行う．

4.5.1 共有可能データと秘匿データ

データ市場の大きなメリットの一つは，秘匿データの顕在化である．一般に，企業や個人が保有するデータは個人の識別性などのプライバシーの問題，そしてビジネス機会の損失のリスクから，公開されることはほとんどない．そのため，データの存在についても知る機会がなく，有益なデータについての情報でさえ入手が難しい状況にある．データ市場によって，今まで個人や企業において保管されてきたデータが市場というプラットフォームに乗り，利用方法が議論可能となる．

世の中には様々な種類のデータが存在しているが，データ市場の登場により，注目されているデータの特徴の一つに，データの共有条件がある．DJ に記述されたデータの共有条件には様々なものが存在するが，大きく 2 つに大別すると，オープンデータなどの一般に共有可能なデータと，売買や交渉により共有される可能性のある秘匿データである．共有可能データは Web 上に公開されていたり，情報開示を求めれば必要に応じて入手可能な状態にあるデータを意味する．一方で，共有できないデータとは，売買や交渉が必要であったり，公開によるリスクを考慮して共有されないデータを指す．McKinsey Global Institute が 2013 年に公開したレポートによれば，図 4-21 に示すように，世の中に存在するデータの中で，オープンデータ以上に秘匿データの方が膨大である（Manyika et al., 2013）．すなわち，秘匿データの方がリソースとしての価値の可能性があるということが言える．

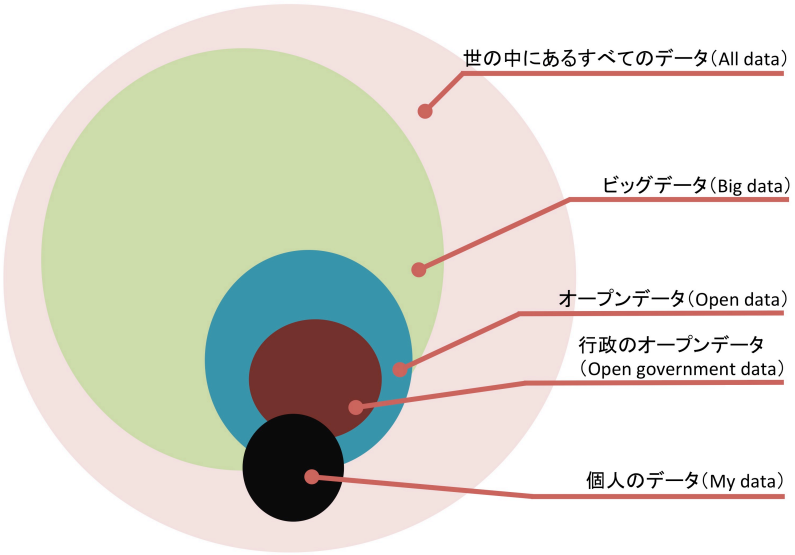

図 4-21　How open data relates to other types of data
（Manyika et al., 2013）（著者により，一部改変）

以上のように，データ市場では，様々なデータやデータに関する情報がやりとりされ，価値を見出され始めている．では，データ市場においてどのようなデータの利用価値が高いと認識されるのであろうか．特に，一般に，「無料のものよりも有料のものの方が質がいい」という先入観が存在する．例えば，街頭に設置されているフリーペーパーよりも，書店でお金を払うことで入手できる情報誌の方が情報量が多く，信頼ができるといった考えである．いくつかの分野において，この考えは成立し，正しい場合がある．データにおいても同様に，無料で入手可能なデータよりも秘匿された有償のデータの方が質が保証されているという先入観があるかもしれない．それでは，実際にデータ市場において，一般に共有可能なデータと共有できないデータのどちらの方に利用価値あると認識されるのだろうか．

前述したように，データの入手以前に利用価値を策定することは困難である．なぜなら，データは適切な文脈が与えられなければ人間の意思決定に役立つ情報あるいは知識にならないからである．また，ビジネス機会の損失のリスクなどから，利用方法が定まっていない状態でのデータ交換は成立しにくい．そして，データ市場において，データとは未だ価値が見出されていない商材であることを考慮すると，データの価値を直接測定するのではなく，データにどの程度利用価値があるのか期待する度合い（利用期待度）を測定することが適当であると考えることができる（Hayashi & Ohsawa, 2016b）．つまり，データ利活用方法検討ワークショップ IMDJ において，要求を満たしたデータ組合せ案（ソリューション）創出に用いられた DJ がデータの利用期待度の高いデータであると見なすことができる．本

節の研究では，DJ のソリューション創出に用いられた回数を利用期待度として測定することで，共有可能データと秘匿データにおけるデータの利用価値について考察する．さらに，本章 4.4 節で開発した DJ 検索システム DJ ストアの DJ 閲覧履歴を用い，検索行動におけるユーザーのデータの利用期待度も測定して比較する．

4.5.2 実験

本実験の目的は，オープンデータなどの一般に共有可能なデータと秘匿されたデータのどちらに価値があると認識されるかについて調べることである．本実験では，データ利活用方法検討ワークショップ IMDJ においてソリューション導出に用いられた DJ を利用価値の高い DJ と見なし，DJ のソリューション創出に利用された回数を共有条件によって比較する．また，本実験では，4.4 節で開発した DJ ストアの閲覧履歴を利用し，ユーザーが閲覧した DJ の数を取得する．DJ ストアの閲覧回数が多い DJ は利用期待度が高いものであるとし，閲覧による利用期待度を共有条件によって比較する．

本研究では，データの共有条件のうち，共有可能データと秘匿データの利用期待度を比較することが目的である．そのため，DJ に記述されている共有条件のうち，「一般に共有可能」であるものを共有可能データとした．また，「条件・交渉が必要」，「範囲を限定して共有可能」，「共有不可能」，「未定」，「購入により共有可」，「研究目的に限って共有可」，「その他の条件付」であるものをまとめ，秘匿データとした．また，DJ の記述は，データ保有者が記入可能な情報のみを記入する形式となっているため，いくつかの DJ には共有条件が記載されていない．そのような DJ は共有条件を判断することができないため，本実験では除外した．

IMDJ における DJ の利用期待度の比較には，DJ ストアに RDF 化して格納されている全 13 回の IMDJ の実施記録データを用いた．各 IMDJ には 10 から 15 人程度が参加し，20 から 30 件程度の DJ が用いられた．IMDJ の実施手順は，本章 4.4 節の DJ ストアの性能実験と同様である．また，DJ の可視化には KeyGraph を用いており，各 IMDJ の実施時間は約 90 分である．

DJ ストアにおいて，自然言語によるクエリで関連する DJ のタイトル及び概要が取得される．DJ のタイトルをクリックすることで，変数ラベルやデータ概要などの詳細情報が取得される．この詳細情報まで取得された DJ を閲覧 DJ として，本実験ではユーザーの興味のある DJ であると見なして数を数える．本実験において DJ ストアには 909 件の DJ，392 件の要求，333 件のソリューションが保存されている（総トリプル数 21,290 トリプル）．また，DJ ストアの DJ 取得方法は 4.4 節の実験と同様に，RDF ストアに sparqlEPCU を用い，DJ ストアにユーザーが入力した文章をキーワードに分割する際に，「データ」や「情報」

などの極端に頻度の高い名詞，「.」や「,」などの記号，助詞・助動詞は不用語として除外した．DJ の閲覧履歴のログデータは，DJ ストアが公開された 2015 年 1 月 31 日から 2016 年 4 月 16 日までのユーザーのアクセスログをサーバーから取得し，ユーザーの検索行動による閲覧された DJ の ID と閲覧数を得た．なお，IMDJ 及び DJ ストアにおいて，データの共有条件についての情報はユーザー及び参加者に提示していない．

4.5.3 結果と考察

DJ ストアから取得した IMDJ における DJ の利用回数から，共有可能データと秘匿データの平均利用回数を得た（表 4-2）．利用回数については，共有可能データと秘匿データには個別の対応がないため，まずは 2 群の分散が等しいと見なせるかどうか f 検定を行った．f 検定の結果，$p < 0.01$となり，等分散が仮定できなかったため，等分散を仮定しない，対応のない t 検定を行った．すると，共有条件が共有可能であるデータの利用回数は 1 件あたり 0.43 回，秘匿データの利用回数は 1 件あたり 1.16 回であり，秘匿データの利用回数が有意に高いことが分かった．

また同様に，DJ ストアから取得した DJ の閲覧履歴から，共有可能データと秘匿データの平均閲覧数を得た（表 4-3）．閲覧数についても利用回数と同様に，共有可能データと秘匿データにおいて個別の対応がないので，2 群の分散が等しいとみなせるかどうかの検定を行った．f 検定を行ったところ，$p < 0.01$となり，等分散が仮定できないため，等分散を仮定しない対応のない t 検定を行った．その結果，共有可能データの閲覧回数は 1.78 回，秘匿データの閲覧回数は 4.39 回となり，秘匿データの閲覧回数の方が有意に高いことが分かった．なお，利用回数と閲覧回数にて登録された DJ 件数が異なるのは，DJ ストアでは非公開 DJ（組織内の利用に限って登録された DJ）は閲覧できない仕様になっているためである．すなわち，利用回数にはある組織内の IMDJ で用いられた非公開 DJ も含んでいるため，閲覧回数と比較して登録数が多くなっている．

表 4-2　利用回数の共有条件による比較

	DJ 件数	IMDJ における平均利用回数
共有可能データ	495	0.430±1.604
秘匿データ	343	1.163±2.193
p-value		**

**: p<0.01, *: p<0.05, $n.s.$: 有意差なし

表 4-3　閲覧回数の共有条件による比較

	DJ 件数	DJ ストアの平均閲覧回数
共有可能データ	486	1.776±4.470
秘匿データ	247	4.393±6.366
p-value		**

**: p<0.01, *: p<0.05, $n.s.$: 有意差なし

以上の結果より，実際のデータ利活用方法検討の場におけるデータの利用期待度の比較において，共有可能データよりも秘匿データの方がおよそ 3 倍多く利用されていることが分かり，データ利活用における意思決定者は，秘匿データの方に利用期待度を高く示していることが分かった．また，検索による閲覧回数を比較したところ，秘匿データの方が約 2.5 倍多く閲覧されており，ユーザーは秘匿データの方に興味がある可能性が示唆される．

注目すべきなのは，利用回数及び閲覧回数の両方において，秘匿データの件数は共有可能データと比較して登録 DJ 件数が少なかったにもかかわらず，秘匿データの利用回数及び閲覧回数が有意に高かったということである．また，興味深いのは，この実験では DJ に記載されているデータの共有条件をユーザー及び IMDJ 参加者に提示していなかったことである．つまり，ユーザー及び IMDJ 参加者は秘匿データであることを事前に知らずに，自身の興味のあるデータを閲覧，あるいは利用価値の高いデータであるとしてソリューション創出に用いているということになる．

この結果の要因の一つとして，秘匿データの目新しさがあると考えられる．例えば，秘匿データには，「大手ホテル予約サイトの履歴」や「来院患者のレセプト」など，オープンデータとして公開され得ないデータについての情報が含まれている．このようなデータの存在は DJ 化されなければ存在を知り得なかったデータである．一方，共有可能データには，「鯖江市の米の生産量推移」や「気温・降水量の長期変化傾向」など，オープンデータ化しているものや API などで入手可能なデータとなっている．このようなデータは Web 上で入手可能なものが多く，DJ 化されていなくても存在を知り得るデータとなっている．つまり，秘匿データはその存在が顕在化していないデータであるため，IMDJ の参加者及び DJ ストアのユーザーにとって目新しさを有しており，利用に対する期待の度合いが高くなり，利用回数及び閲覧回数が共有可能データと比べて有意に高くなったものと考えられる．

以上の研究結果より，一般に共有が困難なデータほど，利活用方法を提案する上で有益なデータであると考えられる可能性が高いということが分かった．つまり，オープン化できないデータほど，提案者及び利用者にとって問題解決及び新ビジネス創出において有用

性が認められる可能性が高いということである．さらに，データに関する情報を検索するユーザーの検索行動においても，一般に共有不可能なデータの閲覧数の方が高いという傾向から，共有が困難なデータの方がユーザーの興味・関心の度合いも高くなる可能性があることが分かった．

また，データの共有条件について別の興味深い事例が報告されている．4.2 節で述べたように，IMDJ は，データの利活用方法が検討され，データの利用価値を策定するプロセスを有している．つまり，データ利用者の要求を満たしたソリューションを構成する DJ から，データの利用価値が認められたり，データに適用する新しい分析ツールが提案されるなどの成果が得られている．例えば，第 6 章にて説明する「街路灯のデータ」と「Google Maps の地図データ」を組み合わせたソリューションでは，実際に行政との交渉の末，非公開のデータを入手した事例が報告されている（池上ら, 2014）．また，高経年化原子力システムの安全性を議論する IMDJ において創出されたソリューションを見たデータ保有者は, 該当データを共有するための有益な情報（所有者や入手方法など）を新たに提供する傾向が見られたという報告がある（大澤, 2014b）．以上の報告をから，データ利用者及び提案者の一般にオープンにされてないデータ利活用への期待，そしてデータ保有者は自身の保有するデータ利用方法を知りたいというニーズが存在する可能性が高いことが示唆される．IMDJ によるデータ利活用方法の提案により，データ保有者が自身のデータの利活用方法を認識すれば，積極的なデータの交換または売買が行われ，データ市場の活性化の可能性があると言えるだろう．

本実験では，共有条件を共有可能データと秘匿データの 2 つに分けて比較を行ったが，共有可能データの中にも利用期待度が高く，注目されやすいデータが存在する．また，秘匿データにおいても，一部の DJ の利用回数及び閲覧回数が高いなど，同様の傾向が見られることが分かった．本実験の結果を踏まえ，共有条件だけでなく，共有可能データ及び秘匿データの詳細な特徴の違いについて検証していくことが必要であるだろう．

4.6 本章のまとめ

本章では，データ市場における DDI に貢献することを目的とし，データ利活用知識構造化と検索システムによる人間のデータ利活用シナリオ生成支援手法の提案，そして実験的データ市場における参加者のプロセスの観察を行った．

4.1 節及び 4.2 節では，本研究にて用いる基礎技術として，データの概要情報であるデータジャケット（DJ），データ利活用方法検討ワークショップ Innovators Marketplace on Data Jackets（IMDJ）について概説した．4.3 節及び 4.4 節では，ユーザーのデータ利活用方法検討を支援するためのデータ記述モデルとデータジャケットの構造化について検討し，データ利活用知識の構造化と検索システムについて論じた．データジャケットだけでなく，過去のデータ利活用方法検討ワークショップ IMDJ において議論された要求，データ利活用案によって価値を認められたデータの関係をデータ利活用知識としてモデル化し，Data Jacket Store（DJ ストア）を実装した．データ利活用知識構造化により，ユーザーが自分と異なる視点を持つ過去のユーザーが考案したデータの使い道を発見したり，過去の別の人が考案したデータ結合案に注目することによって役に立つデータジャケットを探し出すことが可能であることが評価実験により示された．本研究の DJ ストアと Resource Finder（第 5 章 5.3 節にて説明）を組合せたようなアプリケーションとして，Kandogan et al.（2015）の LabBook，Bhardwaj et al.（2015）の DataHub が存在する．これらのアプリケーションはデータ，人などのメタデータを統合したデータ分析プラットフォームであるが，LOD 関連研究と同様に，データ統合とデータ共有を前提とした協働に関する議論に留まっており，一般的に共有できない秘匿データは対象としていない．また，彼らの研究では，異なる粒度の情報を結合するためにメタデータを用いており，セキュリティ上のリスク低減を目的とした DJ におけるメタデータ化とは異なる．

さらに，4.5 節のデータの共有条件に着目した実験では，一般的に公開不可能な秘匿データに対する利用者の期待の高さを確認した．すなわち，データ市場はオープンデータに代表される公開可能データのみに閉じられた場ではなく，公開が難しい個人や企業のデータ及びその保有者を巻き込むイノベーションの場として機能し得ることが分かった．従来研究では，どのようなデータに価値が認められるのかということについて明らかとなっていなかった．

膨大なデータから必要な知識を発見することが困難であるように，データ市場において複数の領域にまたがって存在するデータ，ステークホルダー，ツールなどすべての要素を考慮することは難しい．それ故，意思決定者の異なる価値観や関心を持つ多様な背景知識，意図に対応して適切に設計し，構造化された知識ベースとそれを検索するシステムが必要となる．本章では，データ市場における「データ」，「ソリューション」，「要求」の 3 つの

要素に着目し，構造化と検索システムについて議論し，過去に検討された知識及びシナリオの再利用は新たな知識獲得に有用であることを示した．そして，秘匿データのほうが共有可能データと比較して利用期待度が高くなる傾向から，データ市場はオープンデータに代表される公開可能データのみに閉じられた場ではなく，公開が難しい個人や企業のデータ及びその保有者を巻き込むイノベーションの場として機能し得ることが分かった．

第 5 章　シナリオ構造化による行動計画立案支援

第 2 章にて説明したように，従来研究では，データ利活用に関わるステークホルダーやリソースなどの要素について検討する意思決定者のプロセスにおいて，どのような要素がどのような過程を経て決定されるのか，ということについて明らかとなっていなかった．そこで，第 4 章では，データ利活用知識を構造化し，再利用する仕組みを提案することで，データ概要情報の検索とデータの組合せを立案する意思決定者を支援できることを実験により示した．しかし，データの組合せのみならず，データ利活用に関わる様々な要素の組み合わせ，すなわち行動計画としてのシナリオの生成が重要であると考えられる．そこで，本章では，データ利活用に関わる諸要素が，データ利活用方法を議論する意思決定者の行動，すなわち意思決定者の行動計画立案プロセスにどのように現れるのかを実験的データ市場における意思決定者の行動から観察し，生成されたシナリオの構造化について議論する．

本章は 5 つの節で構成される．5.1 節では，本章の核となる基礎技術である，データ利活用シナリオ創出手法アクション・プランニング（AP）について説明する．続いて 5.2 節では，データのみならず，データ利活用に関わるステークホルダーやリソースなどの要素について考察する AP のプロセスにおいて，どのような要素がどのような過程を経て検討されるのかについて実験的に考察する．5.3 節では，AP によって生成されたシナリオを構造化し，再利用する仕組みににについて論じる．そして，データ利活用による新規事業創出のためのシナリオ生成支援手法として，ステークホルダー表出と関係推定システム Resource Finder（RF）を実装し，文脈によって異なるステークホルダーのシナリオへの関係を推定可能であることを実験的に評価する．続いて，5.4 節では，シナリオ生成支援手法の一つとして，データ市場に新たにデータを取得したい人がどのような変数を取得することが，意思決定に役立つのかという情報は蓄積されてこなかった問題に対して，データ概要から変数ラベル推定方法について議論し，その性能を評価する．そして，5.5 節では本章のまとめを行う．

5.1 アクション・プランニング（Action Planning）

5.1.1 シナリオ

2014 年度に経済産業省のデータ駆動型（ドリブン）イノベーション創出戦略協議会で実施された事業の報告書（経済産業省, 2015）では，データによる新事業創出の障壁として，データの提供及び交渉における時間・労力・合意の不確実性が挙げられている．つまり，異なる領域のデータを入手し，組織間連携を取ってビジネスを興すには，データ提供の方法や合意形成に多大な労力がかかる．確かに，計算機上に保存されているデータは複製が

容易であり，個人を識別可能な情報を含んでいる可能性もある．そのため，ビジネス機会の損失の観点だけでなく，プライバシーやデータ管理コストの観点からも，企業はデータ共有に関して非常に慎重であると言える．つまり，データによる既存のビジネスの付加価値向上や新ビジネスの創出に対する潜在的な期待が高まっているものの，ビジネスとして実際の行動に繋げていくためには，データ分析などの利活用案の創出だけでは不十分である．これらの障壁を乗り越え，データによる既存のビジネスの付加価値向上や新ビジネスの創出という目標を達成するためには，データ分析だけでなく，関連する様々な要素（コスト，ステークホルダー，潜在的リスクなど）を考慮した事業計画及び分析計画といった実行動における指針となるシナリオの生成が必要である．

本研究におけるシナリオとは，データ，経験知，知識から導かれた情報を元に，将来起こり得る事象を論理的に系列化したものである．系列とは，一定の順序に従って並べられた物事のまとまりを意味する．すなわち，事実や要素が互いに矛盾なく繋がれたものがシナリオであり，意思決定者は，このシナリオを読み解くことで，意思決定を行う．しかし，シナリオ創出に膨大な時間とコストをかけて一つの結論を導き出すことは現実的ではない．変化の激しい現代においては単一のシナリオによる戦略はリスクが大きい．政権の転換や法律の改正，大きな事件による社会情勢の劇的な変化などによって，組織は既存の戦略で対応できなくなると，社会の変化に合わせてシナリオを再構築し，新たな戦略を立案しなければならない．特に，長期的な計画であれば早急な対応や計画の修正が必要となる．「戦略寿命の短命化（西村, 2005)」として指摘されているように，刻々と変化する外部環境に対応したシナリオと戦略の創出が重要なのである．

しかし，人間が認知できるのは世界のごく一部であるため，起こり得るすべての可能性を考慮してシナリオを創出し戦略を立案することがほとんど不可能である．つまり，人間の扱う知識は完全なものではなく曖昧なものもあり，知識の完全性を前提とした合理的な判断は期待できない（Simon, 1955）．また，一般化されたフレーム問題（松原・山本, 1987）で指摘されているように，人間も計算機も環境に関するすべての事象を観察し，データを記述することはできない．以上を踏まえると，すべての要素を考慮して意思決定を行うのではなく，不確実性や不完全性を認めた上で意思決定を行う限定合理性に基づくシナリオの生成及び戦略立案が重要であると考えるべきである．

データを活用した新ビジネスを創出したり，データ分析の結果から得られる知見を元に新しい事業を創始するためのシナリオ生成も同様である．データにどのような文脈を与えることで，人間の意思決定において重要な情報あるいは知識とするのかを考える際に，データの組合せを検討するだけでは十分ではない．つまり，実験的データ市場においてデータに文脈を付与するプロセスでは，データの組合せだけでなく，ステークホルダーやリソ

ースといったデータに関わる諸要素の関連性を考慮した検討，すなわちデータ利活用シナリオの生成が重要であると考えられる．

5.1.2 アクション・プランニング概説

他事業者とデータ共有を通して業務連携するためには，データ共有のリスクを極限まで低減しなければならない．リスクを低減するためには，様々な前提条件や関連要素を考慮して策定された行動のシナリオ，すなわちデータ利活用シナリオが必要となる．さらに，意思決定の判断材料となる情報の入手，分析などのすべての行動には時間的・金銭的コストがかかる．そこで，分野横断的なデータ利活用の達成には，データの入手方法，仮説の検証方法と適用する分析方法を含んだシナリオを事前に立案し，検討・評価することが重要となる．

アクション･プランニング（AP）は前章で説明したデータ利活用検討ワークショップ IMDJ において創出されたデータ利活用案（ソリューション）を元に，実行動を促すデータ利活用シナリオ（戦略的シナリオ及び分析シナリオ）を生成するワークショップ手法である（早矢仕・大澤, 2013; Hayashi & Ohsawa, 2013; Hayashi & Ohsawa, 2015b など）．AP では，IMDJ において創出されたソリューションを実行する上で必要な要素の関係性やリスクを論理的に導き出すことで，意思決定を行う際に生じる盲点を低減させ，実行可能なシナリオの策定を行う思考と議論のフレームワークを提供している．様々なステークホルダーが共存するデータ市場において，論理は異なる領域の知識を共通の土俵で議論するための文法である．データ市場は，異なる領域に存在する様々な意図によって取得された多種多様なデータについてステークホルダーが議論する場であることを考慮すると，専門知識が多岐に渡っており，議論の収束は容易ではない．彼らが一堂に会してデータの活用方法について議論し，データの価値を決定するためには，論理という思考のフレームワークが重要となる．計画段階で様々な領域の知識同士の整合していない矛盾を表出化し，前提を疑う議論を促進させることが，盲点となる事実への気付きを促し，実行動におけるリスクを低減させる．

AP では，IMDJ にて創出されたソリューションを実現する際のステークホルダーや必要なリソース（分析技術，人的資源，時間配分，資金配分等）を論理的に検討していくシナリオ生成プロセスを有しており，以下の 3 つのステップを設定している．

1. 要求分析：IMDJ で創出されたデータ利活用案（ソリューション）から，消費者の要求について考察する．顕在的な要求から，論理的に要求の背景を検討し，潜在的なニーズを導く．
2. 要素表出化：ステークホルダー，競合性，実現コスト，データ，実現までの時間や必要なリソースなどの関係性からソリューション実現に関連する要素を導出する．

3. 要素系列化：要素表出化で表出した要素を系列化する．要素同士の関連性を時系列や因果関係で結合することで，欠けていた要素の存在を明らかにする．

以上の 3 つのステップにより，IMDJ で創出したソリューションをシナリオとして精緻化するのが AP の基本的なプロセスである．

AP はデータ利活用を行う関係者が集い，シートへ記入しながら議論を展開することで進行する．AP シートにはいくつかの項目があり，これらが参加者の議論の方向性と思考の枠組みにおいて制約を与えている．制約は一般的に，創造性や問題解決を阻害すると考えられがちだが，思考の枠組みを制限することは創造的発見を促進させることが知られている（Cropley, 1967; Finke et al., 1996）．AP においては，シート上の項目，異なる背景知識を持つデータ利活用に関わるステークホルダー間のコミュニケーションにおける知識の衝突を新たな探索と考察を促す制約として位置付けている．さらに AP では参加者に時間の制限を課し，シートの記入というアウトプットを参加者全員の共通目標とすることで議論の内容の生産性を向上させることを意図して設計している．

特に，AP は，IMDJ と同様にデータ市場における様々なステークホルダー間のコミュニケーションによる参加者同士の相互作用を重視している．Miyake（1986）は 2 者における協働問題解決の過程において，理解していない側の人間の批判が，理解している人間の理解を促進し得ることを示した．また，石井・三輪（2001）は，創造活動においてアイデアに対する評価活動（肯定，中立，批判）の中で，特に肯定と批判を行うペアのアイデアが高く評価される傾向にあることを報告している．さらに，Miwa（2004）は協同問題解決において異なる戦略を有するエージェントの相互作用によって解の発見が改善されることをシミュレーション及び被験者実験の両方で示した．データ市場における自身の立場，職業，役割の視点を議論に持ち込むことにより，データ利活用という共通の目標を持ちながらも，異なる意図と視点，背景知識，意図を持った参加者同士の議論による非同調的な対話が AP において実現することが期待される．以上のプロセスを経て，ステークホルダー間で合意が形成される．つまり，コミュニケーションによる知識の獲得と視座の交換により，個人では知り得なかったデータの使い方や，問題に対する新しい気付きを得ることができると考えられる．

また，AP では 3 つのステップがあると述べたが，IMDJ と同様に，固定されたルールがあるわけではなく，利用目的や実施形態に合わせてシートの内容，人数や時間などは柔軟に変更してよい．本研究では，データ利活用に関わる諸要素を検討する戦略的シナリオの生成やデータに含まれる変数の組合せを検討する分析シナリオの生成では，利用目的に応じてシートの形態を変更している．

5.2 データ利活用シナリオ生成プロセスの観察

データによる既存のビジネスの付加価値向上や新ビジネスの創出に対する潜在的な期待が高まっているものの，ビジネスとして実際の行動に繋げるためには，データ分析だけでなく，関連する様々な要素（コスト，ステークホルダー，潜在的リスクなど）を考慮した事業計画の作成が必要であると言える．

本節では，データのみならず，データ利活用に関わるステークホルダーやリソースなどの要素について考察する意思決定者のプロセスにおいて，どのような要素がどのような過程を経て検討されるのかについて考察するため，シナリオを生成する被験者の行動を観察し，検討を行う．事業計画立案時の筆記行動に着目し，比較として集団と個人のシナリオ生成プロセスを観察することによって，データ利活用に関わる要素の導出過程とプランニングにおける矛盾の解消行動について考察を行う．

5.2.1 論理に基づく問題解決と矛盾解消行動

大澤らは様々な企業における実験により，データなどの客観的な事実から論理的にシナリオを導くことが問題解決を促進させることを示した（大澤, 2003）．また, Kushiro et al.（2014）は Toulmin の議論モデルを用いた問題発見ツールを開発し, 動作音から真空ポンプの新たな故障要因を論理的に導出した．先行研究では，論理によるアプローチは問題解決において有効に作用することの事例が提示されている．すなわち，事業計画作成などの実社会の問題解決においては，解決すべき問題とその解決方法に関連する要素（実現に必要な技術や関係するステークホルダーなど）を単純に追加するのではなく，論理的に関連要素を導出し解を導く精緻化が重要であると考えられる．

現実の問題解決においては，ある知識ベースから導かれた正しい結論でも，のちの新しい知識によって否定されるということが起こり得る．この原因として，現実世界における推論は，完全な知識だけでなく不完全な知識を仮定して推論を進めることで解に至る仮説推論のプロセスを有しているからと考えられている．つまり，ある知識やデータから結論を導いたとしても，さらに別の知識やデータの追加によって，以前の結論に矛盾が生じ，論理の撤回もしくは修正が必要となる場合がある．これらは論理の非単調性（non-monotonic）と呼ばれており，新しい事実の追加に対する結論集合の増加が非単調であることに由来している．既存の知識では完全な解を導けない問題では，仮説を立てて推論を進め，矛盾なく問題を解決できれば，立てた仮説は正しいと考える仮説推論が行われることが知られている（McDermott & Doyle, 1980 など）．仮説推論では，推論過程で矛盾が生じた場合，今までの仮定を棄却・修正するという行動が現れることが分かっている（Ikeda et al., 1993）．

シナリオ生成手法 AP のシナリオ生成過程は，まだ問題解決方法が存在しない問題に対す

る解決方法を提案し，それを実現するための行動に関わる諸要素を論理的に追加してくことで仮定を立てて推論を進め，シナリオを創出するという点で仮説推論であると考えることができる．AP のシナリオ生成プロセスが仮説推論であるとすれば，シナリオ生成プロセスにおいて矛盾が生じた場合，今までの仮定を棄却・修正するという矛盾解消行動が観察されるだろうと考えられる．非単調な推論過程では，矛盾を解消するために今までの仮定を棄却・修正するために，以前に検討していた仮定まで遡って議論を行わなければならない．計算機においてはバックトラックと呼ばれる手戻りが発生し，計算時間が膨大となってしまうことが知られている．また，人間の製品設計プロセスにおいても手戻りは重大なリスクとして認識されており，開発初期段階でコストや技術リスクを考慮しないような不適切な設計プロセスによって，生産における矛盾や衝突が生じる可能性があることが指摘されている（古賀・青山, 2010）．以上のように，設計プロセスにおける手戻りは重要な課題であり，実ビジネスの事業計画のプランニングにおいても生じ得る問題であると言える．

5.2.2 筆記行動によるシナリオ生成プロセスの追跡

矛盾が生じる要因は，不完全な知識を用いた推論過程を経ていることによる非単調性であると考えられる．開発初期の設計や事業計画は後の生産プロセスや実行動に多大な影響を与え得るため，AP においてもデータ利活用方法検討段階におけるシナリオ内の要素の無矛盾化は重要な課題である．そのため，シナリオ生成プロセスにおける矛盾の発生を推定することが求められる．しかし，矛盾箇所を推定するために，すべての被験者のシナリオ生成プロセスにおける思考過程をヒアリングすることは現実的ではない．なぜなら，実験者の介入によりシナリオ生成の議論や進行が阻害されてしまう可能性があるからである．本研究では，ある推論を試行し，矛盾が生じたらバックトラックを行うという被験者の行動に注目し，筆記行動からシナリオ生成における矛盾を解消する行動を追跡する手法を実施する．問題解決に取り組む中で筆記行動を追跡し，解を導くプロセスについて検討した研究はいくつか行われてきた．しかし，Cox & Brna（1995）や山崎・三輪（2001）の研究では，図形・絵などの図的な筆記プロセスが主な対象であり，本研究の対象は筆記活動から論理的な問題解決と矛盾解消行動の推定を行う点で異なっている．

また，AP は集団における思考活動からシナリオを生成するモデルを導入しているが，個人のシナリオ生成プロセスにおいても，問題解決における論理を構築する上で，自身のシナリオの一貫性及び根拠が欠如，知識の矛盾があることに気付き，自身の仮説を部分的に棄却することがある可能性が考えられる．そこで，本研究では比較のため，集団と個人の両方に AP を実施してもらい，筆記行動から矛盾を解消する行動を追跡し，観察を行う．

5.2.3 実験

本実験の目的は，データに文脈を与えるプロセス，すなわち，データ利活用に関わるステークホルダーやリソースなどの要素について検討する意思決定者のプロセスにおいて，どのような要素がどのような過程を経て決定されるのか，ということをシナリオ生成時の被験者の行動を観察することで理解することである．前述したように，データ利活用シナリオ生成プロセスには非単調性が現れる可能性が考えられるが，先行研究では明らかとなっていない．また，データ市場では，意思決定者は様々なステークホルダーの意図を理解・解釈しながらコミュニケーションを通して事業計画を立てる．そこで，他のステークホルダーとのインタラクションを含んだ集団の事業計画立案のプロセスと，他のステークホルダーとの相互作用がない個人の事業計画立案のプロセスを比較して観察することで，シナリオ生成プロセスについて理解することができると考えられる．

本実験では，具体的に以下の 3 つの事項について実験により考察する．

① 個人のシナリオ生成プロセスにおいて，矛盾を解消する行動（要素の削除，要素の追加など）が行われるか．

② 集団のシナリオ生成プロセスにおいて，矛盾を解消する行動（要素の削除，要素の追加など）が行われるか．

③ 以上の①及び②の結果を踏まえ，個人と集団のシナリオ生成プロセスの違いを明確にし，シナリオ創出というデータに文脈を与える人間のプロセスについて考察する．

本実験では，被験者として 20 歳から 40 歳の学生及び社会人 51 人を集め AP を実施した．AP は，データ利活用検討ワークショップ IMDJ において創出されたアイデアをシナリオとして精緻にするプロセスであるため，事前課題として約 90 分の IMDJ を行い，AP に用いるアイデアを創出してもらった．事前課題の IMDJ におけるテーマは，「安全・安心に暮らせる社会を作るためにデータを用いたソリューションを考案する」である．

続いて，個人と集団のシナリオ生成プロセスを比較するため，IMDJ を実施した後に，個人を 11 人，集団を 12 のグループ（1 グループは 3〜4 名）に分け，それぞれ AP で用いるアイデアを選択してもらった．被験者のシナリオ生成プロセスを追跡するため，筆記の軌跡をデータとして取得可能なデジタルペンデバイス（日立マクセル製アノト方式デジタルペン DP-201）を用いた．AP におけるシナリオ生成には，図 5-1 に示すシートを用いた．

本実験の AP では，5.1 節で説明した「要求分析部」を省き，「要素表出化部」及び「要素系列化部」のみを用い，被験者にはシートへの記述を文章またはキーワードで行うよう教示した．なぜなら，要求分析部にはステークホルダーの潜在要求や社会的要因について考察する記入欄が存在するため，筆記行動の追跡により欠けている要素を特定することが難

しいと判断したからである．また，要素系列化においてはアイデア実現までの段階を時系列で記述する形式のものを選んだ．その理由は，図による記述を制限するためである．図は認知的負荷を軽減し，被験者の理解を促進させることを示す研究報告がある（Larkin & Simon, 1987; 諏訪, 1999 など）が，本実験では，シナリオ創出における要素の表出化と系列化という言語化による論理の作用を観察することが目的であるため，敢えて図による記述を制限した．しかし，論理関係を示すための矢印や文字の強調（囲いや下線），取り消し線は許可した．各集団及び個人では始めにアイデアに関連する要素を図 5-1 の要素表出部に記入し，続いて実現までのプロセスを要素系列化部で検討してもらった．被験者には，要素表出化部を検討したのち，要素系列化部に移るように指示を出し，系列化部を記述している段階でシナリオに欠けている要素に気づいたり，再検討が必要となった場合は以前に検討していた部分を加筆または修正することを許可した．また，集団における筆記者はグループ内から選ばれた 1 名のみとした．なお，実験時間は最大 60 分とし，シナリオ生成が完了した時点でデジタルペンによる記入作業を終えるように指示した．

全てのグループでシナリオ記入が完了したのち，各グループまたは個人で創出したシナリオを発表し，被験者同士の相互評価を行った．評価は 5 段階の主観評価（1：とても悪い，2：悪い，3：普通，4：良い，5：とても良い）を導入し，評価軸として従来の AP の評価に用いている新規性（シナリオ内の提案が既存のものにはない新しさを有している度合い），有用性（シナリオ内の提案が問題解決に有効に作用する度合い），実現性（シナリオ内の提案を実現する上で必要な要素が検討されている度合い）を採用した（Hayashi & Ohsawa, 2013; 2015）．

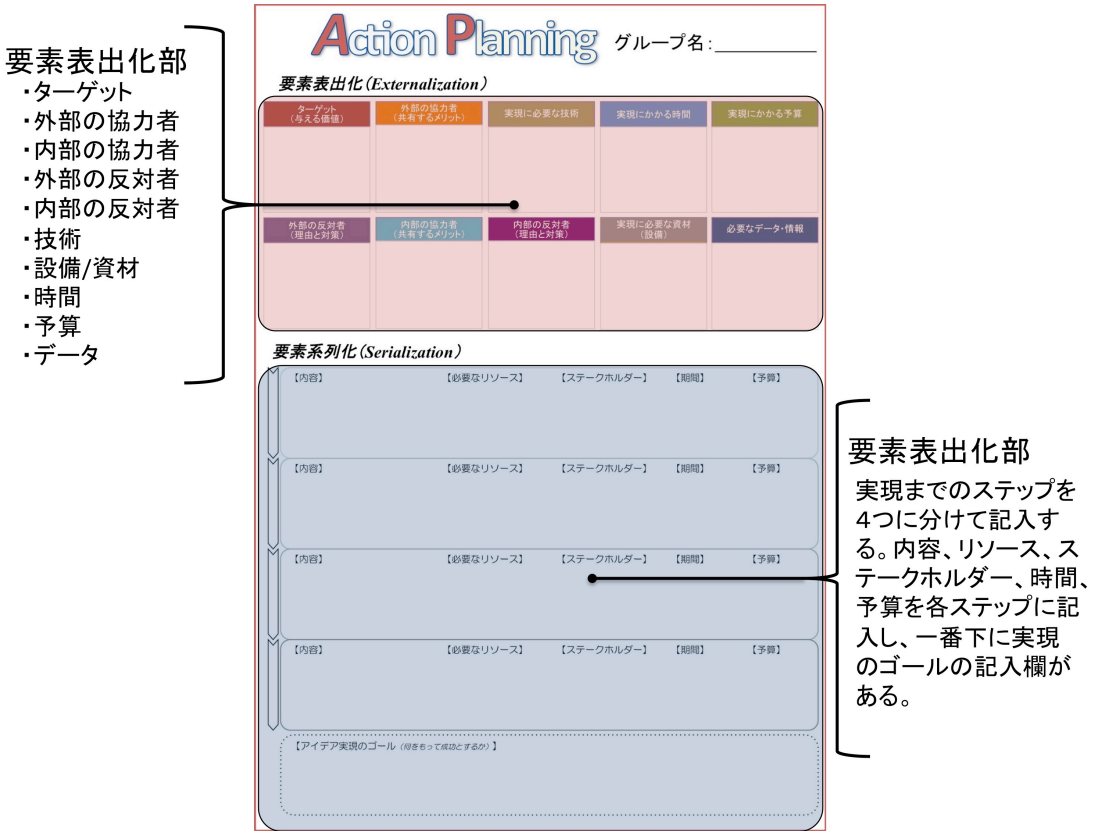

図 5-1　実験に用いたアクション・プランニングシート

5.2.4 結果と考察

5.2.4.1 矛盾解消行動の比較

始めにデジタルペンを用いて被験者の AP 作業中の筆記データを取得し，筆記軌跡の可視化を行った．図 5-2 はある集団における AP の筆記軌跡を表した図である．例えば，図中の番号 22 から 23 は要素系列化から表出化に手戻りが発生した部分を指している．また，番号 12 から 13 は内容からリソースの検討が行われた推移を表しており，番号 15 から 16 は，次の段階の内容を検討しているうちに，前回の段階で必要なリソースに気付き，矛盾を解消する行動（ここでは要素の追加）を行っていることを示している．

まず，筆記行動の可視化により，シナリオ生成プロセスにおける矛盾を解消する行動として，要素の「削除」と「追加」が観察された．削除とは以前にシナリオに記入していた要素をシナリオから除外する行動を指し，追加とは以前に検討していた部分に新たに要素を追記する行動を意味する．表 5-1 は集団及び個人における矛盾解消行動回数の比較を表したものである．要素の削除は，AP のシート上で取り消し線やバツ印の記入によって以前に検討していた部分を修正する行動回数から算出した．

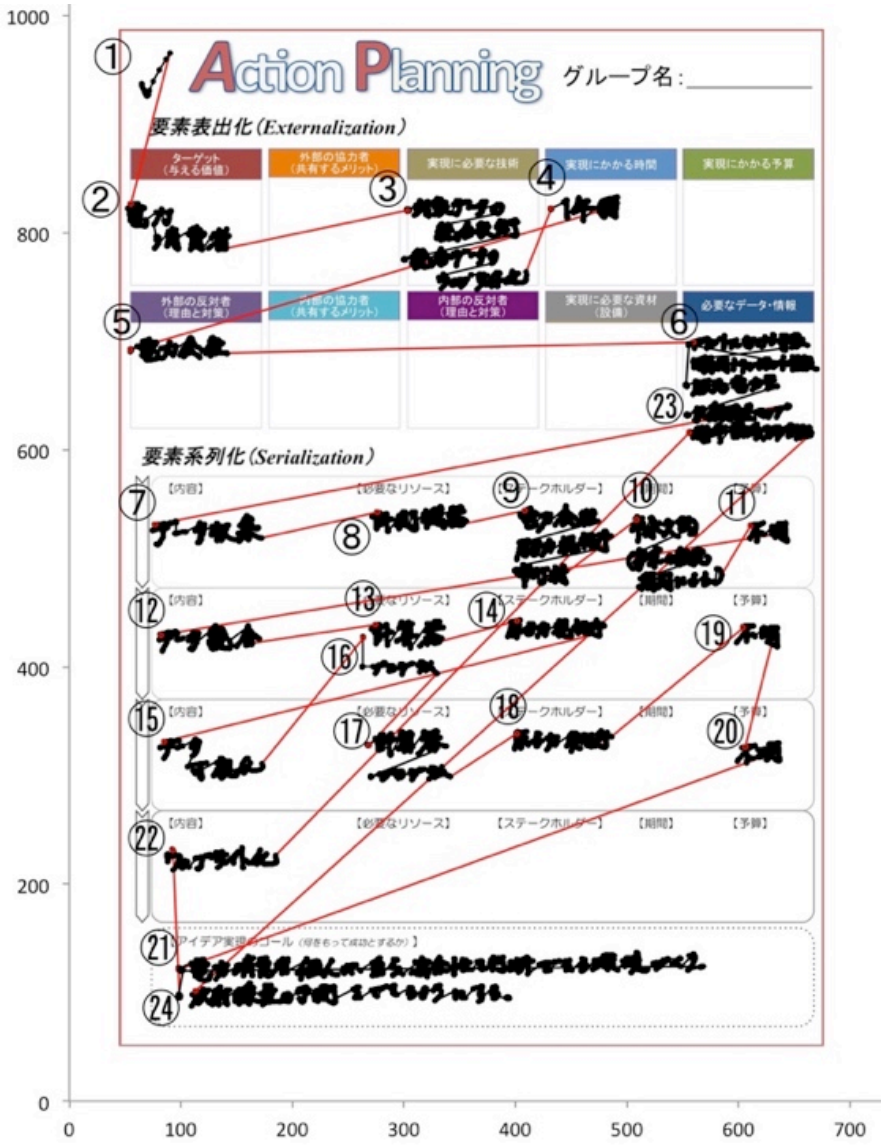

図 5-2　アクション・プランニングへの記述（番号は記述が行われた順番を意味する．また，赤い線は異なる記述項目に移行した部分を表す．①に現れているチェックマークはデジタルペンの仕様であり，シートへの記述が開始されたことを記録するマーカーである．図中の縦軸と横軸は記録された文字情報を描画するための座標を表している．）

続いて，集団と個人のシナリオ生成プロセスを比較するため，対応のない t 検定による分析を実施したところ，集団と比較し，個人によるシナリオ生成の方が要素の削除及び追加行動回数が有意に高いことが分かった（削除：$t(21) = 2.44, p < 0.05$; 追加：$t(21) = 2.84, p < 0.01$）．以降の分析についても集団と個人のシナリオ生成プロセスの比較には対応のない t 検定を用いるものとする．

表 5-1　矛盾解消行動の平均回数（平均値±標準偏差）

	要素の削除	要素の追加
集団	0.250 ± 0.595	1.667 ± 1.491
個人	2.000 ± 2.045	7.000 ± 5.257
p値	*	**

*: $p < 0.05$,　**: $p < 0.01$,　$n.s.$: 有意差なし

5.2.4.2 推移回数の比較

前項では，集団と個人のシナリオ生成プロセスには相違点があり，特に矛盾解消行動に違いがある可能性が高いことが分かった．続いて，集団と個人のシナリオ生成プロセスにおいて，シートへの記入行動に違いが見られるかどうか比較を行う．記入行動として，APのシートにおける要素表出化部と系列化部の移動に着目し，シナリオ生成過程における記入部分の推移回数を比較した．図 5-3 は集団と個人における推移回数の平均を図示したものである．表出化部内推移とは，AP のシートにおいて表出化部内の項目間の移動を表す．また，系列化部内推移とは，シート上の系列化部内の項目間の移動を表している．この結果から，表出化部から系列化部への推移回数と，系列化部から表出化部への推移回数の両方に，有意な差が見られた（$t(21) = 3.08, p < 0.01$; $t(21) = 3.88, p < 0.01$）．一方，表出化部内の推移回数及び系列化部内の推移回数においては有意な差は見られなかった（$t(21) = 0.71, p = 0.48$; $t(21) = 0.74, p = 0.47$）．

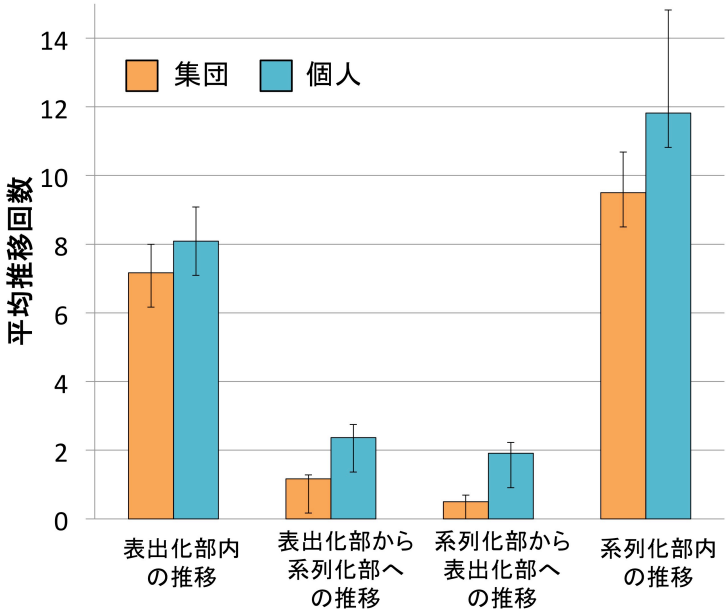

図 5-3　シナリオ生成プロセスにおける推移回数比較

5.2.4.3 矛盾解消行動箇所の比較

前節の分析により，矛盾解消行動及び表出化部と系列化部間の推移頻度が集団と個人のシナリオ生成プロセスの相違点である可能性が高いことが分かった．続いて，AP シート上のどの部分で矛盾解消行動が起こるのかということについて，集団と個人のシナリオ生成プロセスにおける違いを調べた．

表 5-2 及び表 5-3 は，矛盾解消行動（要素の削除及び追加）が起こった箇所を集団と個人で比較したものである．個人におけるシナリオ生成プロセスでは，要素の削除と追加の両方において，特に系列化部から表出化部への推移時に矛盾解消行動が有意に起こっていることが分かる．

さらに集団内，個人内での推移の矛盾解消行動回数に差があるかどうか調べた．集団と個人の各推移における矛盾解消行動回数について分散分析を行ったところ，集団の推移箇所では表出化部内の推移において矛盾解消行動が有意に多いことが分かった（$F(3,44) = 3.55, p < 0.05$）．一方，個人においては推移箇所に互いに有意な差はないことが分かった（$F(3,41) = 1.64, n.s.$）．

表 5-2　要素の削除が起こった箇所の平均回数（平均値±標準偏差）

	表出化部内推移時	表出化部から系列化部への推移時	系列化部から表出化部への推移時	系列化部内推移時
集団	0.167 ± 0.552	0.000 ± 0.000	0.000 ± 0.000	0.083 ± 0.276
個人	0.364 ± 0.643	0.182 ± 0.386	0.727 ± 0.750	0.727 ± 2.004
p値	$n.s.$	$n.s.$	**	$n.s.$

*: $p < 0.05$,　**: $p < 0.01$,　$n.s.$: 有意差なし

表 5-3　要素の追加が起こった箇所の平均回数（平均値±標準偏差）

	表出化部内推移時	表出化部から系列化部への推移時	系列化部から表出化部への推移時	系列化部内推移時
集団	1.083 ± 1.187	0.083 ± 0.277	0.167 ± 0.373	0.333 ± 0.471
個人	2.000 ± 1.348	0.455 ± 0.656	1.364 ± 0.979	3.182 ± 4.468
p値	$n.s.$	$n.s.$	**	*

*: $p < 0.05$,　**: $p < 0.01$,　$n.s.$: 有意差なし

5.2.4.4 思考時間と記入時間の比較

続いて，集団と個人のシナリオ生成プロセスにおける思考時間，記入時間及び総作業時間を比較する．思考時間（TT）及び記入時間（WT）の算出は，デジタルペンを用いた Ikegami & Ohsawa（2014）の実験手法に倣い，5 秒以上の記入作業の停止を思考時間と仮定する式(5.2.4.4.1)と式(5.2.4.4.2)を採用した．tはデジタルペンで取得した時刻を表す．なお，総作業時間は思考時間と記入時間の和で算出するものとした．

$$TT = \sum (t_i - t_{i-1})\ (if\ t_i - t_{i-1} > 5\ seconds)\ (i \in \mathbb{N}) \qquad (5.2.4.4.1)$$

$$WT = \sum (t_i - t_{i-1})\ (if\ t_i - t_{i-1} < 5\ seconds)\ (i \in \mathbb{N}) \qquad (5.2.4.4.2)$$

まず，集団と個人における思考時間，記入時間及び総作業時間の比較を行ったところ，集団のシナリオ創出開始から終了までの総作業時間平均は 1,339 秒，個人の総作業時間平均 2,102 秒であり，個人の作業時間の方が有意に長いことが分かった（$t(21) = 2.75, p < 0.05$）．同様に，思考時間，記入時間ともに個人の方が有意に長い（$t(21) = 2.42, p < 0.05$; $t(21) = 2.07, p < 0.05$）．しかし，AP シートのある項目への記入が行われてから次の項目への記入が行われる間の思考時間をttと表記するとすると，集団のtt平均は 48.54 秒，個人のtt平均は 52.90 秒であり，有意な差は認められなかった（$t(21) = 0.44, p = 0.66$）．この結果は，集団と個人のシナリオ生成プロセスおいて，シート上の各項目にかける思考時間及び記入時間の長さに差はなく，集団の総作業時間が有意に短いことを意味している．

さらに，AP シートの要素表出化部 10 項目，要素系列化部 21 項目，合計 31 項目がどれほど記入されたか（記入率）を集団と個人において比較を行った（図 5-4）．表出化部記入率，系列化部記入率，全体記入率ともに集団と個人においてほとんど差が見られなかった．

続いて，集団と個人のシナリオの評価値について比較を行った．表 5-4 は集団と個人において生成したシナリオの各評価軸（新規性，有用性，実現性）の 5 段階評価の平均である．集団と個人のシナリオ評価では，集団の方がやや高いものの，有意な差は見られなかった（新規性：$t(21) = 1.16, p = 0.25$; 有用性：$t(21) = 0.45, p = 0.65$; 実現性：$t(21) = 0.71, p = 0.48$）．

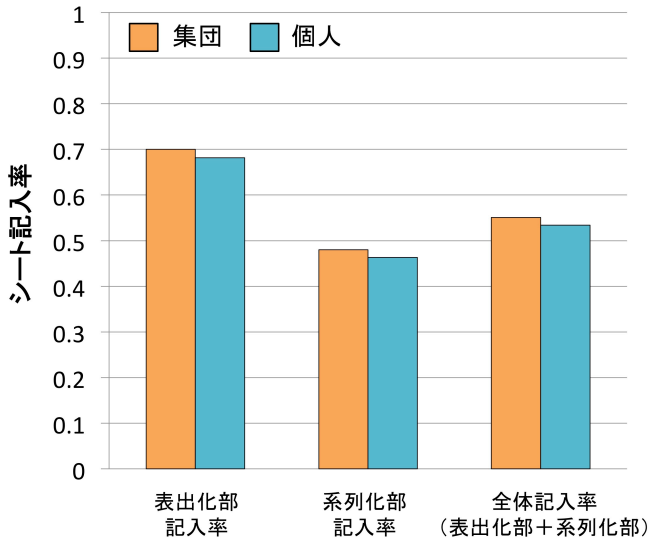

図 5-4　シート記入率比較

表 5-4　シナリオ評価平均比較（平均値±標準偏差）

	新規性	有用性	実現性
集団	3.795±0.468	4.122±0.303	3.783±0.410
個人	3.580±0.365	4.071±0.176	3.682±0.194
p値	$n.s.$	$n.s.$	$n.s.$

*: $p < 0.05$,　**: $p < 0.01$,　$n.s.$: 有意差なし

5.2.4.5 考察

集団と個人の両方において，以前に検討していた部分の記述内容を削除または新たな要素を追加することで矛盾を解消する行動が観察されたことから，AP におけるシナリオ生成では，被験者が非単調な推論過程を経ている可能性が高いことが示唆される．さらに，集団と個人のシナリオ生成プロセスにおいて，要素の削除及び追加行動として現れる矛盾解消行動は，個人において特に多く見られた．

シート内の推移回数を比較すると，集団と比較して個人のシナリオ生成プロセスでは表出化部と系列化部間の移動が多く，記入行動にも違いが見られた．さらにこの結果を踏まえ，矛盾解消行動が起こる箇所を比較したところ，集団では表出化部内の推移において矛盾を解消する行動が多く見られ，個人ではプロセス全体で矛盾解消行動が起こり得ることが分かった．そして，集団と比較して個人のシナリオ生成プロセスでは系列化部から表出化部への推移時の矛盾解消行動が多く行われることが観察された．以上の結果，集団では

要素表出化部で矛盾解消行動が多いことから，シナリオ生成初期の段階で矛盾の解消が行われることで，系列化部における矛盾の発生を低減させている可能性があると考えられる．

そして，AP シートへの記入時間，思考時間及び総作業時間の比較では，集団と個人においてシート上の各項目にかける思考時間及び記入時間の長さに差はなく，集団の総作業時間が有意に短いことが確認された．さらに作業終了後のシートの記入率は集団と個人でほとんど差がなく，両者のシナリオの評価値（新規性，有用性，実現性）は同程度であった．以上のことから，集団と比較し時間的なコスト（1.57 倍の時間）と手戻りコスト（4.30 倍の矛盾解消行動）を考慮すれば，個人においても集団において生成したシナリオと同程度の評価及び完成度のシナリオを生成可能であることが分かる．

そして，個人及び集団のシナリオ生成プロセスにおいて，要素の追加や削除という矛盾を解消する行動が観察され，AP におけるシナリオ生成プロセスには仮説推論における非単調性が現れることが分かった．さらに，集団と個人のシナリオ生成プロセスには明確な違いがあることが検証された．集団による思考よりも個人における思考の方が頻繁に矛盾が生じる可能性があることが分かった．さらに，集団におけるシナリオ生成では，シナリオ生成初期（表出化部）において矛盾の解消が多く行われていたことから，系列化部における矛盾の発生を低減させている可能性があると考えられる．また，集団と個人のシナリオ生成において，矛盾解消行動による手戻りコストや思考時間，記入時間の短縮性を考慮すると，集団の方がシナリオ生成作業効率としては高い．しかし，手戻りコストの発生とシナリオ生成時間の延長を認めれば，個人においても集団において生成したシナリオと同程度の評価及び完成度のシナリオを生成することが可能であることが分かった．

以上をまとめると，本節の実験では，AP を用いたシナリオ生成プロセスにおいて，要素の追加や削除という矛盾を解消する行動が観察され，実験的データ市場におけるシナリオ生成プロセスには，仮説推論における非単調性が現れることが分かった．また，実験的データ市場におけるシナリオ生成，すなわちデータに文脈を付与するプロセスでは，データの組合せだけでなく，ステークホルダーやリソースといったデータに関わる諸要素の関連性を考慮した検討プロセスが重要であるという示唆が得られた．

本節では，データのみならず，データ利活用に関わるステークホルダーやリソースなどの要素について考察する意思決定者のプロセスにおいて，どのような要素がどのような過程を経て検討されるのかについて考察するため，シナリオを生成する被験者の行動を観察した．実験的データ市場においてデータに文脈を付与するシナリオ生成プロセスには仮説推論における非単調性が現れることが分かった．また，データの組合せだけでなく，ステークホルダーやリソースといったデータ利活用に関わる諸要素の関連性を考慮した検討が重要であるという示唆が得られた．

5.3 シナリオ構造化によるステークホルダー表出と関係推定

5.3.1 シナリオにおけるステークホルダー

シナリオとは，データや知識から導かれた情報を元に，将来起こり得る事象を系列化したものである．意思決定者は創出されたシナリオを読み解き，理解することで実行動におけるリスクを低減することができる．本研究におけるシナリオの創出は，すでに知識として確立した行動を策定するのではなく，まだ知識として確立していない行動，特にデータ利活用による新事業の創出におけるシナリオを対象としている．

ステークホルダーという語は「利害関係者」という意味で用いられることが多いが，本研究では Freeman（1984）の定義である「組織の目的達成に影響するまたは影響を受ける集団あるいは個人」に基づき，シナリオ実現に関与する集団または個人と定義する．第 3 章で述べたように，近年，異なる分野のデータ交換や保有データの利活用に対する期待は高まっているものの，データ利活用による新事業創出を目的としたビジネスでは，データの分析，欲しいデータの入手方法は知識として確立していない．さらに，データ分析の結果などを実際のビジネスに活かすためには，各段階で様々な障壁が存在しており，特に，自社内を含むデータ利活用に関わる様々なステークホルダーの検討の重要性が指摘されている（経済産業省, 2015）．しかし，膨大な情報から問題解決に必要な知識を発見することが困難であるように（Simon, 1955），複数の領域にまたがって存在する全てのステークホルダーを考慮することは難しい．Ward & Chapman（2008）は，ステークホルダーはビジネスまたはプロジェクトにおいて主要な不確実性を生む主要因であると述べ，文脈及び関連するステークホルダーを明確にする事の重要性を主張している．

また，ビジネス環境の複雑化に伴い，マネジメントする対象のステークホルダーの複雑性は増している．そのため，意思決定者は異なる価値観や関心を持つ多様なステークホルダーとの関係，つまり事業シナリオへの関わり方を明確にすることが必要であると考えられる．しかし，ステークホルダーの関わり方は，ビジネスの文脈によって異なる．本研究における文脈とは，シナリオの内容及びシナリオが実現される状況を意味する．例えば，「街路灯情報を地図上にマッピングし，夜間に明るいルートを提案するアプリケーションの開発」という事業シナリオがあるとする．この文脈において，住民はシナリオ実現に協力的な立場を取るであろう．しかし，アプリケーションによって暗い地域が明らかとなり，そのエリアの夜間の犯罪発生件数が増加するという懸念から，夜間に暗い地域に生活する住民はシナリオ実現に反対する可能性が考えらえる．一方，「街路灯情報から夜間に暗い地域を明らかにし，新たな街路灯設置エリアを発見する」という事業シナリオの場合，暗い地域の住民は協力的な立場になると考えられる．しかし，新しい街路灯設置と維持のコストから，行政関係者が反対するステークホルダーとして新たに表出するかもしれない．この

ように，同じステークホルダーでも事業シナリオによって，ステークホルダーのシナリオへの関わり方は異なる可能性がある．つまり，ステークホルダーのシナリオへの関係はシナリオの文脈に依存する（図 5-5）．

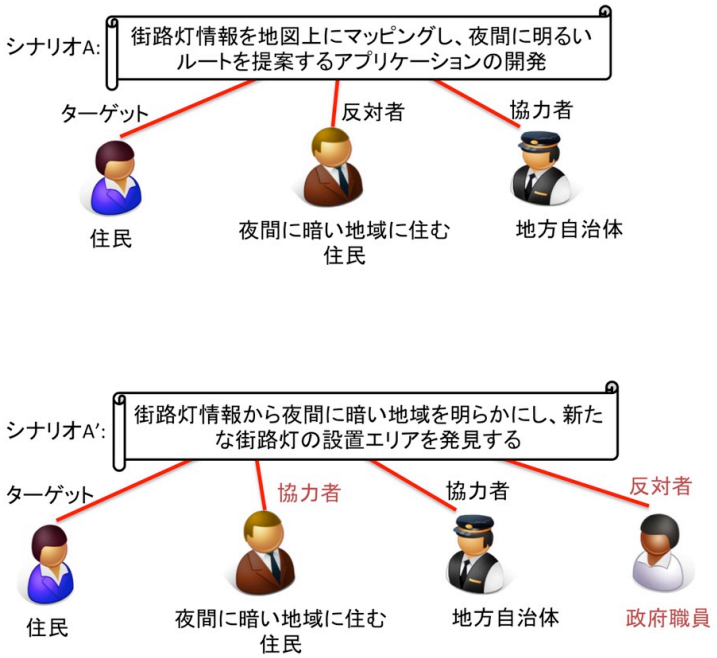

図 5-5　シナリオ及びステークホルダーとその関係の例

以上のように，新事業創出において，ステークホルダーが誰であり，シナリオにどのように関わってくるのかということは，シナリオ実現において検討すべき重要な課題である．しかし，異なる分野を横断したデータ利活用に関わるステークホルダーとその関係についての組み合わせは膨大となり，全てのステークホルダーを考慮して実現し得るシナリオを策定することは困難である．そこで，人間のシナリオ創出の支援として，新事業に関わるステークホルダーを推薦するシステムが必要であると考えられる．

本研究では，データ利活用による新事業創出という既存の知識では解決が困難な問題に対して，ステークホルダー推薦システムを提案する．知識ベースとして DBpedia（Auer, 2007 など）とデータ利活用シナリオの RDF ストア（Hayashi & Ohsawa, 2015a; 2015c）を用い，シナリオに関連すると考えられるステークホルダーについての情報を表出化し，シナリオ生成時に気付きえない潜在的に関係のあるステークホルダーを推薦し，シナリオにおける関係を推定するシステムとして，Resource Finder（RF）を実装する．

5.3.2 Resource Finder の設計

5.3.2.1 シナリオの構造化と再利用

第 4 章 4.4.節 DJ ストアの開発により，過去に行われた議論や問題解決に用いられ価値が認められたデータについての情報が IMDJ の参加者だけに留まらず, オンライン上で参照可能となった．データ市場の活性化のためには，過去のデータに基づくソリューションやシナリオによって価値の判断されたデータやユースケースを一般に入手可能な状態にすることも重要である．ユーザーはワークショップに参加していなくても，過去のデータ利活用に関する情報を引き出すことにより，データの価値やシナリオに必要な知識を得ることが可能になる. 本研究では実行動を促すシナリオ生成手法 AP によって創出された戦略的シナリオを知識ベース化し，過去にシナリオに用いられた知識（シナリオに関わるステークホルダー，実現に必要なリソース，データなど）を再利用する仕組みを提案する．これにより，シナリオに欠けている知識要素を推定し，推薦するシステムが構築可能となる.

AP によって創出される戦略的シナリオの要素表出化部では，シナリオに関わるステークホルダー（ターゲット，外部の協力者，内部の協力者，外部の反対者，内部の反対者）及びリソース（データ，技術，時間，資材，予算）が記入される．これらの要素を RDF で記述することで，シナリオにおける知識要素の役割と関係性を表現することが可能となる.

例として,「明るいルートを検索できるアプリケーション」というソリューションを実現するシナリオに関わるステークホルダーとリソースを RDF で記述したものを図 5-6 に示す. 例えば,「明るいルート検索アプリケーションというシナリオのステークホルダーで，ターゲットとしているのは住民と行政である」という情報は，戦略的シナリオを表す主語を“scenario:明るいルート検索アプリケーション”，矢印で表記される述語を“ap:stakeholder”と“ap:target”，目的語を“行政”と“住民”とした RDF による有向グラフ形式で表現することができる．シナリオに含まれるリソースについても同様の表現により記述可能となる.

RDF で記述した複数の戦略的シナリオをデータベースに格納し，知識要素との関係性を述語によって繋げることで，シナリオに欠けている知識要素を推定し，推薦するシステムを構築することが可能となる．例えば，ターゲットが共通するシナリオ同士は，他のステークホルダーも共通している可能性が高いと考えられる．すると，一方のシナリオに含まれているステークホルダーから，他方に欠けているステークホルダーを推定し，シナリオを作成する意思決定者に推薦することが可能となる.

図 5-7 は A 社，B 社が考案したシナリオ（X，Y）とそれぞれのシナリオに関わるステークホルダーとその関係性を RDF によって表現したものの例である．A 社と B 社のシナリオには，“sh:1”，“sh:2”，“sh:3”が共通のステークホルダーとして含まれており，その関係性についても一致する．すると，A 社と B 社のシナリオはステークホルダーにおいて関係が

強いと考えることができ，A 社のシナリオに含まれているが B 社のシナリオには含まれていない社外の協力者である“sh:4”を B 社のシナリオにも関係すると推定し，B 社の意思決定者に推薦することができる．また，本手法によるシナリオのデータ構造化により，シナリオに含まれるデータやその分析方法などのリソースについても類似するシナリオに含まれる知識要素から推定し，シナリオに欠けている要素を意思決定者に推薦できる．

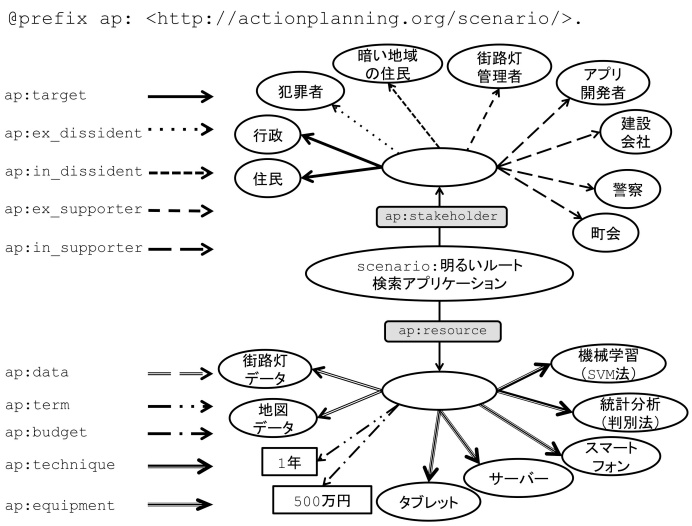

図 5-6　戦略的シナリオの RDF によるグラフ形式表現の例（シナリオに関連するステークホルダー及びリソース部分のみ表示．無地のノードは空白ノードを表す）

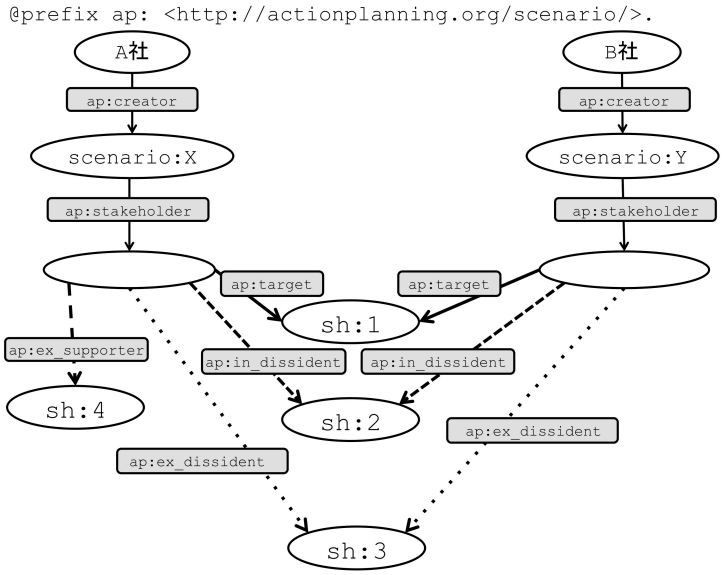

図 5-7　シナリオに関係するステークホルダーを RDF グラフ形式で表現した例（ノードの sh はステークホルダーを意味し，リンクに付いているラベルは述語を表す）

5.3.2.2 ステークホルダー候補のシナリオへの関係性推定

前項では，AP によって検討されたシナリオに含まれる知識要素のデータ構造化と再利用方法について論じ，システムの実装を行った．この推薦システムの実装により，過去のワークショップで検討されたシナリオに含まれていた知識を再利用し，シナリオ実現に欠けている知識を補完することが可能となる．また，第 1 章にて，本研究におけるシナリオの創出は，すでに知識として確立した行動を策定するだけではなく，まだ知識として確立していない行動，特にデータ利活用による新事業の創出におけるシナリオを対象としていることを述べた．前述したように，複数の領域にまたがって存在する全てのステークホルダーを考慮することは難しい．

今までに人間の発言や行動から人間関係を推定する研究（神田・石黒, 2000 など）や，論文の共著関係から人間の繋がりを可視化する研究（松尾ら, 2005 など）は行われてきた．しかし，文脈に応じて変わり得るステークホルダーの事業への関わり方についての研究は十分議論されてきたとは言い難い．そこで，本研究では，関係性が未知のステークホルダーが与えられた時に，シナリオにどのような関係となりうるのかを推定する仕組みが求められる．なお，本研究における関係とは，シナリオとステークホルダーの関係であり，ステークホルダー間の関係を意味しているのではない．前述したたように，ステークホルダー及びそのシナリオへの関わり方は，シナリオ自体の文脈に依存する．つまり，同じステークホルダーであっても，シナリオによって協力者や反対者になり得るということである．

以上の議論を踏まえ，シナリオを$scenario$，ステークホルダーを$stakeholder$，関係を表す述語を$relationship$と表現すると，あるシナリオにおけるステークホルダーの関係は述語で$relationship(scenario, stakeholder)$と記述することができる．述語である$relationship$としてターゲット（$target$）や協力者（$supporter$）を定義すれば，シナリオにおけるステークホルダーの様々な関わり方を記述することができる．

5.3.2.3 関係推定のための学習データ作成

続いて，前項にて構造化し RDF で記述した AP によるシナリオを構成する要素のうち，ステークホルダー（述語が“ap:stakeholder”）で連結した要素群とその関係を用い，訓練データを作成する．シナリオの文脈把握には，各シナリオに含まれる単語の出現頻度や重要語の抽出などが必要である．これらを求めるために，Latent Dirichlet Allocation（Blei et al., 2003）などのトピックモデルの手法を用い，文書のトピックを抽出する方法が考えられるが，AP によって創出されるデータ利活用シナリオは様々な領域の問題を扱っており，シナリオ内に含まれる単語は高次元かつスパースである．そのため，トピックを識別するのに十分な情報が含まれておらず，学習が収束しない恐れがある．また，異なる領域の問題を

扱うシナリオの場合，一つのシナリオが複数のトピックを有することが考えられる．さらに，シナリオの種類によってトピックとなり得る主要部分の位置は様々である．

そこで本研究では，文脈をシナリオに含まれる単語の集合と仮定して計算を行うこととした．AP のシナリオによる訓練データではシナリオに含まれる単語 1 つが独立して 1 つの文脈を表す仮定する．例えば，前述した「街路灯情報から夜間に暗い地域を明らかにし，新たな街路灯設置エリアを発見する」というシナリオの場合，「街路灯」，「夜間」，「暗い」，「新たな」，「設置」，「エリア」，「発見」という互いに独立した文脈を持つということになる．なお，「情報」や「データ」などの頻出語及び句読点などの記号類はストップワードとして除外している．以上の仮定により，ある文脈におけるステークホルダーとその関係を定めるというアプローチについて考える．

例えば，先ほどの例で言うと，「街路灯」の文脈でステークホルダーとして「住民」がターゲット，「地方自治体」が協力者であるとする．これらの文脈におけるステークホルダー及びその関係が学習されていれば，他のシナリオにおいて「街路灯」という文脈が与えられていたとき，「住民」及び「地方自治体」がそれぞれターゲット，協力者であるということが推定可能となる．同様に，「観光」の文脈でステークホルダーとして「外国人」がターゲット，「ホテル」が協力者という関係にあるとする．ステークホルダー及びその関係が未知の新しいシナリオ（「観光」という語を含む）を入力としたとき，その文脈から「外国人」がターゲット，「ホテル」が協力者であると推定することができる．

以上の手続きにより，AP によって創出されたデータ利活用シナリオ 60 件と，各シナリオに含まれる総ステークホルダー数 679 件とその関係から，文脈・ステークホルダー・関係のデータセットを 51,625 件作成し，学習データとした．また，シナリオにおけるステークホルダーの関係は「ターゲット（target）」，「協力者（supporter）」，「反対者（dissident）」の 3 つを抽出した．なお，通常の AP のシナリオでは「協力者」と「反対者」について，それぞれ内部及び外部の視点を入れてステークホルダーを定めている．しかし，内部及び外部の視点はシナリオの文脈よりも，シナリオ作成者の社会的立場が反映されるため，本研究では「内部の協力者」と「外部の協力者」をまとめて「協力者」，「内部の反対者」と「外部の反対者」をまとめて「反対者」として統一した．

5.3.2.4 関係の計算方法

前項で述べたシナリオを訓練データとし，文脈を表す単語間の独立性を仮定することで，ある文脈が与えられたときのステークホルダーの関係をターゲット，協力者，反対者に分類する問題として扱うことができる．

今，あるシナリオ（$scenario_x$）が与えられたとき，$scenario_x$は文脈（$context_i$）の集合

と見なすことができる．以上を考慮すると，$scenario_x$は式(5.3.2.4.1)と表すことができる．

$$scenario_x = \bigcup_{i=1}^{N} context_i \qquad (5.3.2.4.1)$$

続いて，各文脈$context_i$を用いて訓練データからステークホルダー（$stakeholder$）とその関係（$relationship$）を取得する．あるステークホルダー（$stakeholder_j$）に注目したとき，シナリオ$scenario_x$を構成する各$context_i$における$relationship$（$target$，$supporter$，$dissident$の 3 種類）の出現頻度を計算することで$stakeholder_j$の$scenario_x$における$relationship$を求めることができる．

例えば，$context_i$において$stakeholder_j$が$target$である頻度を求める場合，文脈・ステークホルダー・関係からなる学習データから，$context_i \cap stakeholder_j \cap target$となるデータセットを取得し，その数を数えることで求める．つまり，$context_i$において$stakeholder_j$が$relationship_k$であることを$rel_k(context_i, stakeholder_j)$と表すとすると，$scenario_x$において$stakeholder_j$が$relationship_k$となる回数は式(5.3.2.4.2)と表すことができる．$target$と同様の方法で$supporter$及び$dissident$についても出現回数を計算し，その中で最も出現回数が高い関係が最終的に$scenario_x$における$stakeholder_j$の関係となる．

$$rel_k(scenario_x, stakeholder_j) = \sum_{i=1}^{N} rel_k(context_i, stakeholder_j) \qquad (5.3.2.4.2)$$

例えば，「観光」の文脈におけるステークホルダー「外国人」のターゲット尤度を計算するためには，AP のシナリオを RDF で記述したデータベースから，$(context_i = \text{"観光"}) \cap (stakeholder_j = \text{"外国人"}) \cap (relationsip = \text{"target"})$となるデータを取得する．「観光」の文脈におけるステークホルダーとその関係を RDF ストアから取得する SPARQL クエリは図 5-8 であり，取得されるデータの例は表 5-5 である．

```
SELECT ?rel ?st (COUNT(*) AS ?count) WHERE {
?s rdfs:label ?title;
<http://actionplanning.org/scenario/outline> ?scenario;
<http://actionplanning.org/scenario/id> ?id;
<http://actionplanning.org/stakeholder> ?st_URI.
?st_URI ?rel_URI ?st.
FILTER (contains(?scenario, "観光"))
BIND(
            IF (contains(str(?rel_URI),"http://actionplanning.org/stakeholder/ex_dissident"), "反対者",
               IF (contains(str(?rel_URI),"http://actionplanning.org/stakeholder/target"), "ターゲット",
                  IF (contains(str(?rel_URI),"http://actionplanning.org/stakeholder/in_dissident"), "反対者",
                     IF (contains(str(?rel_URI),"http://actionplanning.org/stakeholder/ex_supporter"), "協力者",
                        IF (contains(str(?rel_URI),"http://actionplanning.org/stakeholder/in_supporter"), "協力者","関係不明")
                        )
                     )
                  )
               ) AS ?rel).
} GROUP BY ?rel ?st ORDER BY ?rel
```

図 5-8　「観光」の文脈におけるステークホルダーとその関係を RDF ストアから取得する SPARQL クエリ

表 5-5　RDF ストアから取得した「観光」の文脈におけるステークホルダーと関係

関係	ステークホルダー	出現回数	関係	ステークホルダー	出現回数
ターゲット	バス会社	1	協力者	店舗（広告主）	1
ターゲット	商店	1	協力者	旅館	1
ターゲット	地方情報誌	1	協力者	日本政府観光局 JNTO	1
ターゲット	外国人観光客	1	協力者	日本観光局	1
ターゲット	小売店	1	協力者	映像解析開発会社	1
ターゲット	旅行会社	1	協力者	商店	1
ターゲット	旅館	1	協力者	自治体の観光課	1
ターゲット	日本人観光客	1	協力者	行政	1
ターゲット	自治体	1	協力者	行政機関	1
ターゲット	航空会社	1	協力者	訪日観光経験者	1
ターゲット	観光地	1	協力者	警備会社	1
ターゲット	鉄道会社	1	協力者	航空会社	2
ターゲット	飲食店	1	協力者	観光協会	2
ターゲット	訪日外国人	2	反対者	一次産業従事者	1
協力者	アプリ開発会社	1	反対者	住民	1
協力者	カメラ開発会社	1	反対者	儲かっていない店舗	1
協力者	グルメ情報サイト	1	反対者	口コミで悪く書かれた人	1
協力者	システム開発部	1	反対者	口コミで悪く書かれた場所	1
協力者	タクシー	1	反対者	地域住民	1
協力者	バス会社	1	反対者	自社保有データの外部提供を拒否する人	1
協力者	データ提供者	1	反対者	観光客	1
協力者	地域情報誌	1	反対者	観光雑誌	1

また, 例えば,「外国人観光客の Twitter のつぶやき情報から隠れた観光スポットを抽出し, 推薦サービスを提供する」というシナリオを入力した際, 図 5-8 に示すクエリを式(5.3.2.4.1)に従ってシナリオ内の文脈に分割してクエリを作成し，RDF ストアに発行する．表 5-6 に

出力例を示す．なお，表中の関係尤度（$likelihood_k$）は式(5.3.2.4.3)から求めた．

$$likelihood_k = \frac{rel_k(scenario_x, stakeholder_j)}{\sum_k rel_k(scenario_x, stakeholder_j)} \quad (5.3.2.4.3)$$

$(k \in \{target, supporter, dissident\})$

表 5-6　「外国人観光客の Twitter のつぶやき情報から隠れた観光スポットを抽出し，推薦サービスを提供する」というシナリオを入力し，AP のデータベースから推薦されたステークホルダーとその関係（関係尤度が最も高い値のものを太字で示す）

ステークホルダー	ターゲット尤度	協力者尤度	反対者尤度
行政	8.11	**62.2**	29.7
航空会社	40.7	**59.3**	0
ホテル	48.9	**51.1**	0
小売店	48.9	**51.1**	0
システム運用部	0	**100**	0
バス会社	**50.0**	**50.0**	0
飲食店	**50.0**	**50.0**	0
鉄道会社	**50.0**	**50.0**	0
自治体	**68.4**	21.1	10.5
訪日外国人	**100**	0	0
データサイエンティスト	0	**100**	0
観光客	21.4	0	**78.6**
アプリ開発者	0	**100**	0
観光協会	16.7	**83.3**	0
旅行会社	**91.7**	8.33	0
口コミデータ提供事業者	0	**100**	0
口コミで悪く書かれた場所	0	0	**100**
地域住民	19.0	0	**81.0**
地域情報誌	47.6	**52.4**	0
日本観光局	0	**100**	0

5.3.2.5 未知のステークホルダーへの対応

以上の方法では，AP において過去に検討されたシナリオに含まれるステークホルダーを用いて，ステークホルダー及びその関係が未知の新しいシナリオが与えられた際に，シナリオの知識ベースからステークホルダー及びその関係を推定し，ユーザーに推薦する仕組みであった．しかし，社会には様々なステークホルダーが存在しており，AP のシナリオで表出したステークホルダーはそのごく一部にすぎない．そのため，データ市場における潜在的なステークホルダー及びその関係の表出が重要となる．

そこで，「あるシナリオが与えられたとき，その文脈においてどのステークホルダーがどのように関連するのか」ということを推定するため，DBpedia Japanese[26]（以降，DBpedia と呼ぶ）の構造化された情報を用いる．DBpedia は Wikipedia の記事を RDF を用いて構造的に記述し，SPARQL エンドポイントから取得できるようにしたサービスである．DBpedia に含まれる構造化されたステークホルダーに関する情報を用いることで，過去に検討されたシナリオ内のステークホルダーだけでなく，シナリオに関係すると推定されるより多くの分野のステークホルダーを抽出できると考えられる．まず，DBpedia からステークホルダー群を抽出した．本研究では，ステークホルダー群として，DBpedia に含まれる職業一覧（<http://ja.dbpedia.org/resource/職業一覧>）に記載のある職業名と，それらの職種からリンク（<http://dbpedia.org/ontology/wikiPageWikiLink>）するリソースの職業名及び概要を取得した．さらに Wikipedia において「職業」のカテゴリに属するリソースについても同様に取得し，合計 633 件のステークホルダー候補を DBpedia の SPARQL エンドポイントから取得した（図 5-9）．

```
select distinct ?name ?outline where {{
            <http://ja.dbpedia.org/resource/職業一覧> <http://dbpedia.org/ontology/wikiPageWikiLink> ?uri.
            BIND (replace (str(?uri), 'http://ja.dbpedia.org/resource/', '') AS ?name)
            OPTIONAL{
                        ?uri rdfs:comment ?outline.
            }
            } UNION {
            <http://ja.dbpedia.org/resource/Category:職業> <http://dbpedia.org/ontology/wikiPageWikiLink> ?uri.
            BIND (replace (str(?uri), 'http://ja.dbpedia.org/resource/', '') AS ?name)
            OPTIONAL{
                        ?uri rdfs:comment ?outline.
            }
}}
```

図 5-9　DBpedia に発行した職業情報を取得する SPARQL クエリ

5.3.2.6 Resource Finder の入出力

本項では，新規シナリオの入力から，関係が未知のステークホルダーとその関係を推定

[26] http://ja.dbpedia.org/

し，ユーザーに結果を提示するまでの Resource Finder（RF）のプロセスについて説明する．事前処理として，DBpedia からステークホルダー情報として，前項で説明した手法により，ステークホルダー候補の職業情報のリストを取得する．続いて，過去の AP にて検討されたシナリオから，ある文脈におけるステークホルダーとその関係を格納した学習データを用意する．以降，過去のシナリオから得られたステークホルダーを$stakeholder_{ap}$，DBpedia から取得したステークホルダーを$stakeholder_{db}$と表記することとする．

① ユーザーはステークホルダー及びその関係が未知の新規事業シナリオ（$scenario_x$）を創出し，RF に自然言語にて入力する．

② シナリオの文章を分かち書きし，文脈の集合を得る（$scenario_x = \bigcup_{i=1}^{N} context_i$）．

③ 学習データから，$context_i$における$stakeholder_{ap_j}$とその関係rel_k（$target$, $supporter$, $dissident$の 3 種類）を取得する（$\sum rel_k\left(context_i, stakeholder_{ap_j}\right)$）．

④ $context_i$における$stakeholder_{ap_j}$を元に，$stakeholder_{ap_j}$が有する語を概要情報に含む$stakeholder_{db_l}$を DBpedia から取得する．これにより，$context_i$における新しいステークホルダー$stakeholder_{db_l}$が得られる．

⑤ $stakeholder_{db_l}$の$context_i$における関係rel_kを算出する．$stakeholder_{db_l}$の$context_i$における関係は，$rel_k(context_i, stakeholder_{db_l}) = \sum_j \sum_i rel_k\left(context_i, stakeholder_{ap_j}\right)$と表せる．

⑥ 各文脈$context_i$において$stakeholder_{db_l}$のターゲット，協力者，反対者の出現頻度（$rel_k(context_i, stakeholder_{db_l})$）を計算する．

⑦ 出現頻度が最大となるrel_kが$scenario_x$における$stakeholder_{db_l}$の関係として出力される（$argmax_k \sum_j \sum_i rel_k\left(context_i, stakeholder_{ap_j}\right)$）．

図 5-10 は以上に説明した入力から出力までの概要を例示したものである．

例えば，ユーザーが新しい事業シナリオとして 5.3.2.4 項で例示した「外国人観光客の Twitter のつぶやき情報から隠れた観光スポットを抽出し，推薦サービスを提供する」を入力し，本提案システム RF によって未知のステークホルダーとその関係を取得した例を表 5-7 に示す．なお，$stakeholder_{db_l}$の関係尤度は式(5.3.2.6.1)を用いて算出した．

$$likelihood_k = \frac{rel_k\left(scenario_x, stakeholder_{db_l}\right)}{\sum_k rel_k\left(scenario_x, stakeholder_{db_l}\right)} \qquad (5.3.2.6.1)$$

$(k \in \{target, supporter, dissident\})$

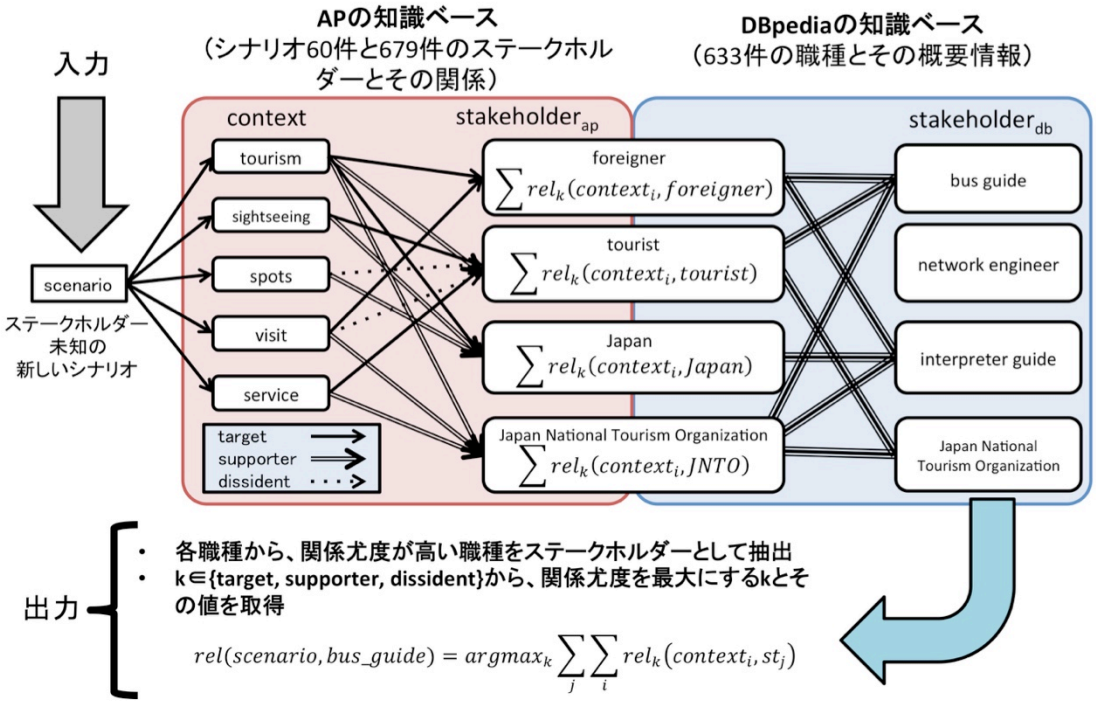

図 5-10　ステークホルダーとその関係が未知の新しいシナリオを入力し，AP の知識ベースから DBpedia の知識ベースに含まれる職業情報を取得し，ステークホルダー及びそのシナリオにおける関係を取得するプロセス

以上に示したように，ステークホルダー及びその関係が未知の新しいシナリオを自然言語にて入力した際に，過去にシナリオ生成手法 AP によって検討されたシナリオの知識ベースだけでなく，DBpedia の知識ベースから未知のステークホルダー及びそのシナリオへの関係を推定する機能を有したアプリケーション Resource Finder（RF）を開発した．

本アプリケーションは DJ ストアと同様に Javascript で実装し，Web アプリケーションとしてオンラインでアクセス可能とした（図 5-11）．

表 5-7 「外国人観光客の Twitter のつぶやき情報から隠れた観光スポットを抽出し，推薦サービスを提供する」シナリオの未知のステークホルダーとその関係の一部（関係尤度が最も高い値のものを太字で示す）

ステークホルダー	ターゲット尤度	協力者尤度	反対者尤度
記者	**40.5**	38.1	21.4
議員	**42.9**	39.64	17.47
ルポライター	**44.4**	41.7	13.9
医療監視員	**59.1**	28.8	12.1
地方公務員	**63.6**	4.55	31.8
落語家	**66.7**	28.56	4.76
旅行作家	**74**	20	6
通訳案内士	35.8	**40.3**	23.89
舞妓	35.8	**44.4**	19.8
海事代理士	31.8	**47.75**	20.5
公共政策コンサルタント	16.5	**47.9**	35.68
ツアーコンダクター	36.7	**48.94**	14.28
航空従事者	18.8	**55.1**	26.05
アイドル	24.5	**60.4**	15.1
IT コーディネータ	5.43	**63.07**	31.53
ネットワークエンジニア	9.73	**64.6**	25.65
観光コンサルタント	23.1	**67.9**	8.959
ウェブデザイナー	14.3	**71.4**	14.3
バイヤー	20.6	**76.3**	3.09
建設コンサルタント	2.91	**93.2**	3.88
電気工事士	15	40	**45**
消防官	15.6	38.29	**46.1**
鳶職	8	36	**56**
中小企業診断士	33.3	8.33	**58.33**
整備士	13.6	27.25	**59.1**
社会保険労務士	14.3	7.14	**78.6**

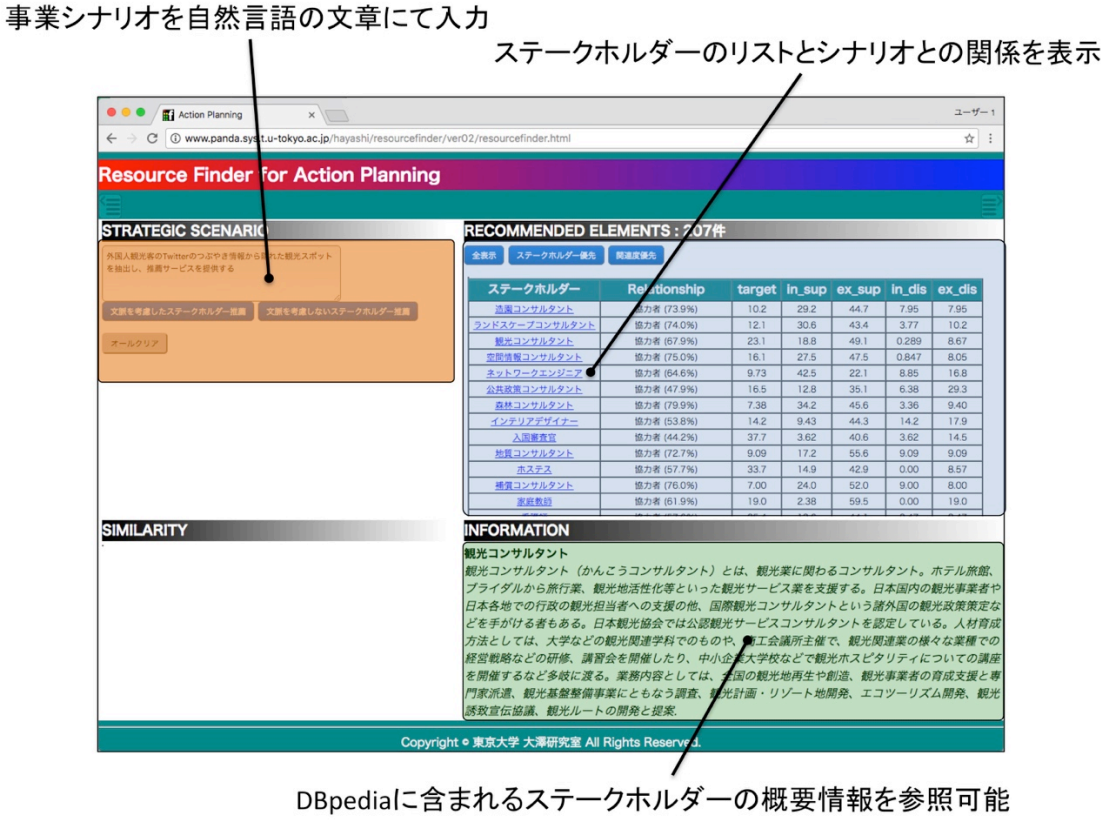

図 5-11　Resource Finder のインタフェース

5.3.3 実験

RF の性能評価を行うための実験を実施した．本実験の目的は，RF によって推薦されたステークホルダー及びその関係の有効性を評価することである．有効性とは，あるシナリオを与えたときに，RF が推定するステークホルダー及びそのシナリオへの関係の候補が意思決定者にとってシナリオに関係があるかどうか，そして，推定されたシナリオへの関係が正確であるかどうかということを意味する．シナリオを RF に入力し，最も関係尤度が高い値を示したステークホルダーをステークホルダー候補とし，職業情報として職業名及び RF で推定されたシナリオへの関係を無作為に 20 件取得した．本実験では，以下に示す 2 件のシナリオ（シナリオ A 及び B）を RF に入力した．

- シナリオ A：外国人観光客に対して Twitter のつぶやき情報から隠れた観光スポットを抽出し，推薦する．
- シナリオ B：個人の医療情報を一元管理する電子カルテを作り，医療の効率化を図る．

なお，RF のアプリケーションに用いた RDF ストアは，DJ ストアと同様に sparqlEPCU を用いた．さらに，入力文章をキーワードに分割する際には TinySegmenter を用い，助詞，助動詞，句読点などの記号類は不用語として除外した．また，RF の推薦結果の性能評価する

に当たり，RF によるステークホルダー候補 20 件に加え，RF の推薦結果に含まれないダミー職種を 20 件無作為に選択した．

各シナリオにおいて合計 40 件の職種が被験者に提示され，①被験者はこれらの候補がシナリオに関係するかの有無を回答する．続いて，②関係があると考えた候補をターゲット，協力者，反対者に分類するという作業を行ってもらった．各シナリオにおいて RF によるステークホルダー候補とダミー職種 40 件をランダムに配置して被験者に見せ，21 人の被験者の回答のうち，過半数の被験者が「関係あり」と回答した職種を「シナリオに関係のある職種」とする．また，過半数の被験者が「関係なし」と回答した職種を「シナリオに関係のない職種」とする．また，詳細な関係（ターゲット，協力者，反対者）においては，RF が提示した関係がどれほど被験者の回答に対して正しいかを評価する．すなわち，RF が提示した関係$k(k \in \{target, supporter, dissident\})$を回答した被験者数を被験者数で割った値の平均をその職種の正答率とする（式(5.3.3.1)）．なお，N_kは関係kとなるステークホルダー候補の数を表す．

$$\text{正答率}_k = \frac{1}{N_k}\sum_{l=1}^{N_k}\left(\frac{(\text{RF の提示した関係を回答した被験者の数})_k}{(\text{被験者の数})} \times 100\right) \qquad (5.3.3.1)$$

また，各シナリオを入力として取得した 20 件のステークホルダー候補は表 5-8 に示す．なお，各シナリオのダミー職種 20 件は以下である．

- シナリオ A：絵本作家，政治家，幼稚園教員，ピアノ調律師，漫画家，歯科医師，声優，マジシャン，納棺師，キックボクサー，軍事評論家，学者，新聞配達員，大学教授，ライトノベル作家，放送作家，シンガーソングライター，牧師，ファッションモデル，不動産屋
- シナリオ B：消防官，選挙屋，潜水士，動物管理官，花屋，花火師，航空事業者，巫女，通訳案内士，弁理士，和菓子職人，振付師，建設コンサルタント，運転手，ノンフィクション作家，マルチタレント，パン屋，ファンタジー作家，酪農家，盲導犬訓練士

表 5-8　RF から取得した各 20 件のステークホルダー候補

シナリオ A		シナリオ B	
ステークホルダー候補	関係（尤度）	ステークホルダー候補	関係（尤度）
医療監視員	ターゲット(59.1)	サラリーマン	ターゲット(57.1)
地方公務員	ターゲット(63.6)	記者	ターゲット(50.8)
旅行作家	協力者（74.0）	社会保険労務士	ターゲット(42.9)
通訳案内士	協力者（40.3）	幼稚園教員	ターゲット(42.9)
公共政策コンサルタント	協力者（47.9）	労働基準監督官	協力者（68.8）
航空従事者	協力者（55.1）	臨床心理士	協力者（66.6）
入国警備官	協力者（57.7）	青年海外協力隊員	協力者（90.8）
ネットワークエンジニア	協力者（64.6）	ネットワークエンジニア	協力者（52.6）
空間コンサルタント	協力者（75.0）	国税専門官	協力者（65.8）
ランドスケープコンサルタント	協力者（74.0）	補償コンサルタント	協力者（80.0）
バイヤー	協力者（76.3）	銀行員	協力者（56.2）
補償コンサルタント	協力者（76.0）	社会福祉士	協力者（74.7）
インテリアデザイナー	協力者（53.7）	IT コーディネータ	協力者（85.1）
建設コンサルタント	協力者（93.2）	国家公務員	協力者（57.1）
農業コンサルタント	協力者（69.3）	国会議員	協力者（71.5）
電気工事士	反対者（45.0）	料理研究家	協力者（75.6）
消防官	反対者（46.1）	Web デザイナー	協力者（73.3）
中小企業診断士	反対者（58.3）	保護監察官	反対者（51.7）
整備士	反対者（59.1）	キャリアコンサルタント	反対者（71.4）
社会保険労務士	反対者（78.6）	事務員	反対者（51.3）

5.3.4 結果と考察

アンケートにて被験者に提示された 40 件の職種のうち，過半数の被験者が「関係あり」としたものを関係のある職種，過半数の過半数の被験者が「関係なし」としたものを関係のない職種としたとき，関係の有無に関する結果は表 5-9 となった．シナリオ A において，RF が「関係あり」と提示した 20 件のステークホルダー候補のうち，15 件が被験者によって関係ありと見なされ，正答率は 75.0%となった．また，RF において「関係なし」，すなわ

ちダミー職種として抽出したものは 20 件中 10 件が関係なしと見なされ，正答率は 50.0%となった．一方，シナリオ B において，RF が「関係あり」と提示した 20 件のステークホルダー候補のうち，18 件が被験者によって関係ありと見なされ，正答率は 90.0%となった．また，RF においてダミー職種として抽出したものは 20 件中 17 件が関係なしと見なされ，正答率は 85.0%となった．以上の結果から，関係の有無に関しては，RF のシナリオに対するステークホルダー候補の関係の有無の推定性能は比較的高いと考えることができる．シナリオ A のダミー職種の正答率が 50.0%とやや低い値となった原因は，政治家，不動産屋，ライトノベル作家，シンガーソングライター，ファッションモデルなどは多くの被験者が関係ありと回答したことにある．これらの職種は RF の推定結果には表出しなかったものの，外国人観光客の日本文化への興味や宿泊業，地方創生などの文脈においてシナリオ A と結びつきやすい．そのため，被験者の多くがこれらの職種を「関係あり」と見なした可能性が考えられる．つまり，観光の文脈は多様なステークホルダーが関連してくる可能性が高く，文脈における制約が比較的弱いと言うことができる．一方で，シナリオ B において関係のない職種の正答率が 85.0%と，シナリオ A と比較して高くなっているのは，医療関係の文脈は制約が強く，関係の有無がある程度明確に分かれていたことが原因であると考えられる．一方，シナリオ B において，消防官，潜水士，動物管理官の職種が関係ありと見なされていたのは，健康，医療，人命救助関連の文脈と結びつきやすいものであったからと考えられる．

表 5-9　職種の関係の有無

	関係のある職種の正答率	関係のない職種の正答率
シナリオ A	75.0%	50.0%
シナリオ B	90.0%	85.0%

続いて，ターゲット，協力者，反対者の詳細な関係について調べたところ，表 5-10 に示した結果となった．シナリオの平均正答率は協力者 43.0%が最も高く，ターゲット 17.1%，反対者 2.89%となった．各シナリオにおいて，RF が協力者であるとした職種については比較的正答率が高かったものの，ターゲット及び反対者の正答率は低い結果となった．

続いて，以上の結果となった要因について考察する．シナリオ A において，RF の推定結果ではターゲットが 2 件，反対者は 5 件あったにも関わらず，ほとんどの被験者はターゲットあるいは反対者を選択しなかった．シナリオ A では，アンケートに含まれる全 40 件の職種において，被験者によって選ばれた詳細な関係の回数は，ターゲットが 58 回，協力者

が 256 回，反対者はわずか 18 回であった．また，シナリオ B においては，ターゲットが選ばれた回数は 149 回，協力者は 143 回，反対者は 11 回と，ターゲットの方が協力者の回数を上回ったものの，反対者が選択された回数は著しく低いことが分かった．この原因として，RF の推定結果において，ターゲット及び反対者が推定結果として表出した回数が非常に少ないことが挙げられる．ターゲット及び反対者の表出回数が少ない要因は，AP の RDF ストアに格納されている 679 件のステークホルダーのうち，ターゲットは 163 件，協力者は 371 件，反対者は 145 件であることが影響している．つまり，協力者はターゲットの 2.23 倍，反対者の 2.55 倍多く存在している．すなわち，シナリオ生成時に意思決定者は協力者を多く表出させていることが分かる．また，その傾向は関係を推定する被験者の行動においても現れており，被験者は各シナリオに対して提示された職種から関係を見出す際に，ほとんど反対者を想定していなかった可能性が考えられる．

本実験の結果，シナリオに対する関係の有無に関しては，比較的高い推定が可能であり，詳細な関係では，協力者については比較的高く推定できる可能性が示唆された．AP による過去に生成されたシナリオの知識ベースの影響もあり，ターゲット及び反対者となり得るステークホルダーの表出は今後の課題である．また，シナリオ生成時に反対者となり得るステークホルダーの表出を促す仕組みとして，提案手法である RF を導入することも検討していく必要があるだろう．

表 5-10　職種の詳細な関係の正答率

	ターゲット	協力者	反対者
シナリオ A	11.8%	54.2%	3.81%
シナリオ B	22.4%	31.9%	1.96%

また，RF の実際のデータ市場における導入事例として，スーパーマーケット事業者における IMDJ 及び AP 顧客の種類と販促アプローチに違いへの気付き促進と輸送分野におけるサービス事業化ワークショップの結果が挙げられる．スーパーマーケット事業者においては，レシピ推薦の有効性の検証のために，必要なステークホルダーへの気付きが促され，店舗アンケートの取得や顧客属性という仮説に到達し，現在も分析プロジェクトが進行している．また，輸送分野においては，気象情報提供会社に関する情報を検索と災害の種類を選定及び道路管理会社へのヒアリングなど，関連するステークホルダーへの気付きから実行動につながっている（経済産業省, 2016）．

本節の研究では，データ利活用による新規事業創出のためのシナリオ生成支援手法として，ステークホルダー表出と関係推定システム Resource Finder を提案した．本手法では，文脈抽出の複雑性を回避するため，シナリオ内の単語の独立性を仮定して文脈を定めた．そのため，「旅行」と「観光」のようにビジネスにおいて結びつきの強いと考えらえる単語を異なる文脈として計算している．今後の課題として，共起性の高い単語を考慮した文脈把握方法を検討する必要があると考えられる．

本システムでは，新たなシナリオを創出する際に，潜在的なビジネスパートナーや必要なデータについての情報を提示することが可能となる．また，知識ベースからの知識の推薦により，敵対する可能性のあるステークホルダーを事前に予期することができる．以上により，計画遂行時のリスク低減に役立てることができると考えられる．

本研究では DBpedia に含まれる職業の概要情報のみを用いてステークホルダー及びそのシナリオへの関係を取得する仕組みを提案したが，DBpedia では様々な情報が階層構造及びネットワーク構造を形成して各リソースが述語によって連結している．今後は職業名だけでなく，さらに細かい職種や具体的な会社名などを取得することでより具体的なシナリオ生成を支援できると考えられる．また，本節ではステークホルダーに着目し，職業のみを推薦対象としたが，データ利活用シナリオにはステークホルダーだけでなく，データ，分析ツール，コストなど他の様々な要素が含まれている．これらの要素についても，本システムを応用し過去のシナリオから学習させることで，ユーザーに推薦する仕組みを構築できると考えている．

5.4 共起性に着目した変数ラベル推定

5.4.1 変数ラベル

変数とはデータの中身であり，データを構成する要素である．例えば，日本の天候データの中には，日付や天候，気温などが含まれている．図 5-12 は 2016 年 3 月における東京の天気の情報という架空のデータと，そのデータをデータジャケットとして記述したものである．例示しているように，実データには変数として「年」,「月」,「日」,「最高気温（℃）」,「最低気温（℃）」,「天候」が含まれており，それぞれの変数に対し，値が格納されている．一方，DJ はがデータの概要情報であるため，データの値を含むのではなく，実データの中身を説明する情報を含んでいる．実データの変数は変数ラベルとして記述され，データの概要情報はタイトル及びデータ概要として表される．変数ラベル（Variable Label）とは，データに含まれる変数の意味または名前である．データ利活用方法の検討では，DJ の変数ラベルの組合せから適用する分析ツールや期待する分析結果に関する議論が行われる．

実データ

年	月	日	最高気温 (℃)	最低気温 (℃)	天候
2016	3	28	16.8	9.0	くもり
2016	3	27	16.1	6.0	晴れ時々くもり
⋮	⋮	⋮	⋮	⋮	⋮
2016	3	2	12.4	1.1	晴れ
2016	3	1	10.3	2.1	快晴

変数（上段）／値（下段）

データジャケット (*DJ*)

タイトル	2016年3月における東京の天気の情報
データ概要 (*OD*)	気象庁が提供している2016年3月の東京における天気のデータ。日々の最高気温、最低気温、そして天候の情報が含まれている。気象庁のWebサイトからダウンロードできる。
変数ラベル (*VL*)	年
	月
	日
	最高気温 (℃)
	最低気温 (℃)
	天候

図 5-12　実データとデータジャケット

天候情報などのデータでは,「年」,「気温」,「天候」といった既知の変数が含まれており，理解することは容易であると考えらえる．しかし，データによって変数及び値の種類は様々であり，それらにはデータの利用目的やデータ取得者の意図，そして思想が大きく反映される．例えば，災害時の避難情報のデータベースを DJ 化したものに含まれる「危険度」と

いう変数は，「災害時の活動困難な度合いを項目ごとに評価した値を考慮した総合危険度」という意味を持っている．また，自動車製造工程に関するデータの DJ には，「乗員への加害性側面衝突試験：横からの衝突に対する乗員への保護性」という専門的な変数が含まれている．DJ では，このようなデータ固有の変数に関して自然言語によって記述された説明文を変数ラベルと定義して用いている．

データ利活用におけるデータ分析とは，データに含まれる変数の値を入力とし，あるルールに従って組み合わせ，変換することによって，出力結果を得るプロセスであるということができる．すなわち，実際のデータ分析の段階ではデータ内の変数の値に対して分析ツールを適用し，値を変換して可視化などの分析結果を得ることとなる．特に第 3 章で述べたように，異なる分野におけるデータ利活用においては，データの中には個人を特定する情報が含まれている場合があり，データ内の変数の値を公開したり，共有することは困難である．そこで，事前にどのようなデータを入手し，どのような分析によってどのような仮説が検証できるのかというシナリオを立て，評価することが重要であると考えられる．

例えば，2016 年 3 月の東京における天候情報に含まれる変数ラベル「日時」,「最高気温」,「最低気温」,「天候」と，あるスーパーマーケットにおけるビールの売上データに含まれる変数ラベル「日時」,「売上」,「個数」,「銘柄」があったとしよう．このとき，共通する変数ラベルである「日時」を基準とし，「最高気温」と「売上」あるいは「個数」を組み合わせることによって，「最高気温の推移とビールの売上は相関関係がある可能性がある」という仮説を立てることができる．つまり，データ自体は秘匿のまま，データの変数ラベルの組み合わせを議論することで，仮説及び検証のための分析プランを立てることが可能となる．すなわち，データ分析の方法を議論する段階においては，データの概要情報として変数ラベルが公開されていることが重要であると考えることができる．

しかし，DJ ではメタデータとしてデータのタイトル，概要説明，変数ラベル，共有条件，データの保存形式などを記述する 12 件の項目を設定しているが，データ保有者にすべての情報を入力することは強制していない．つまり，データ保有者がデータ市場に公開したい情報のみが DJ に記述されるため，登録された DJ には必ずしも変数ラベルが含まれているとは限らないのである．そのため，変数に関する情報の不足により，本来結合する可能性のある DJ 同士が未結合となってしまうという問題が生じ得ると考えられる．

さらに，2016 年 3 月の段階で DJ ストアに登録されている 909 件（非公開 DJ を含む）の DJ のうち，変数ラベルを公開している DJ 数は 763 件，変数ラベルを公開していない DJ は 146 件であった（図 5-13）．つまり，データ自体は秘匿のまま，データ利活用においては変数ラベルの組み合わせを議論することで，仮説及び検証のための分析シナリオを立てることができるが，変数ラベルを公開していない DJ については，変数に関する情報の不足によ

り，データ同士の組合せ可能性を議論することが難しくなる．

そこで，本研究では，DJ に含まれるデータの概要説明の情報であるデータ概要（Outline of Data：OD）と変数に関する情報である変数ラベル（Variable Label：VL）を用い，変数ラベルを含まない DJ に含まれる可能性の高い変数ラベルを推定する方法を提案する．本研究により，変数名や変数に関する情報である変数ラベルが DJ 含まれていなくても，データ概要から変数ラベルを推定することで，データ同士の潜在的な結合可能性を意思決定者に提示し，データ利活用を促すことが可能となると考えられる．

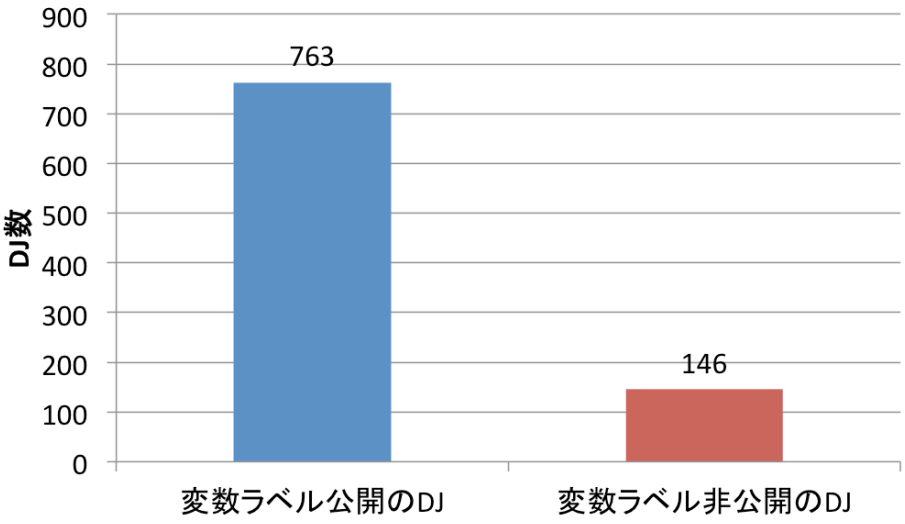

図 5-13　変数ラベルが公開された DJ と非公開の DJ 数比較

5.4.2 変数ラベル推定のアプローチ

本研究の目的は，変数ラベルが未知のデータの変数ラベルを推定することである．しかし，データ本体は秘匿であるため，データ自体を閲覧することによって含まれている変数ラベルを知ることはできない．そこで，DJ に記述されたデータの概要情報を用いることによって，目的を達成するというアプローチを取る．まず，以下の 2 つの仮定を設定し，モデル化することを考える．

- データ本体を説明する情報（データ概要）が似ているデータ同士は類似している
- 同じ変数ラベルの集合を有するデータ同士は類似している

すなわち，データ概要が似ているデータは同じ変数ラベルを有していると仮定する方法である．この仮定によるモデルを以下，モデル 1 と名付けて用いる．

しかし，モデル 1 に表されるデータ概要の類似度だけでは，記述量の多いデータ概要のみが類似度が高いデータとして頻出し，それらに含まれる特定の VL のみが推定結果に現れ

てしまう可能性が考えられる．また，異なる領域に保存されているデータの DJ には，データ取得者の癖やデータの取得意図が反映されるため，データによっては変数ラベルが欠損している可能性がある．そこで，変数ラベルの共起性に着目したモデルを考える．

変数ラベルの共起性とは，同じデータ内に登場する頻度の高い変数ラベルが存在するという特徴である．例えば，位置情報を表すデータであれば，変数ラベルである「緯度」,「経度」は同時に登場する頻度が高い．また，人間の身体情報のデータであれば，変数ラベルである「身長」,「体重」,「年齢」,「性別」は同時に取得される割合が高い変数ラベルの集合であると言える．このように，データ内で同時に登場する頻度の高い変数ラベルの集合が存在するという共起性を用い，データ概要の類似度だけでは補えない欠損している変数ラベルを補完することができると考えられる．この仮定によるモデルを以下，モデル 2 と名付けて用いる．

本研究によって，変数ラベルが非公開の DJ であっても，含まれる可能性の高い変数ラベルを推定し，データ同士の潜在的な結合可能性を意思決定者に提示し，データ利活用を促すことができると考えらえる．さらに，従来の手法では，新たにデータを取得したい人がどのような変数を取得することが，意思決定に役立つのかという情報は蓄積されてこなかった．データの取得には多大なコストがかかるため，もし，データ取得後に取るべきであった変数が明らかとなったとき，新たにコストをかけて取得しなければならないという問題が生じ得る．本手法により，新たにデータを取得し，意思決定に役立てたいと考えるデータ取得者に対し，過去に蓄積された DJ の情報を用いることで，どのような変数を取得することが意思決定において重要であるのかという知見を示すことができる．

5.4.3 データの特徴モデル化の検討

5.4.3.1 データ概要

第 4 章 4.1 節にて説明したように，DJ では，変数ラベルの他にデータについて説明するための情報として，データ概要，共有条件やフォーマット，データの種類など 12 項目が含まれている. 本研究では, データの類似度を図る指標として, DJ に含まれるデータ概要（OD）を用いる方法を提案する．データ概要とは，データがどのようなデータであるのかという説明文を意味する. データ概要を用いる理由は, データ概要は DJ の中で最も記述量が多く，データの特徴を表し，データの内容を理解するのに適当であると考えられるからである．図 5-14 に本実験で用いる 799 件の DJ の各項目の記入率を表す．記入率は，799 件の DJ において，その項目が記入されている割合を表す．タイトルは DJ 登録時の必須記入項目であるため, 記入率は 100%となっている. 次いで最も高い記入率であるのがデータ概要 (95%) である．共有条件，データの種類，保存形式，データの所在が 80%以上の記入率となり，

データの類似度を図る指標の候補になり得ると考えられるが，共有条件，データの種類，保存形式は自然言語ではなく，選択式の記入項目である．また，データの所在は URL などが記入されている．そのため，共有条件，データの種類，保存形式，データの所在は，データの内容を表す特徴として十分とは言えない．以上の理由から，データの類似度を図る指標としてデータ概要を採用する．しかし，5%の DJ にはデータ概要が記述されていない．そのため，本研究では，データ概要にタイトルを加えたものをデータ概要として用いることとする．

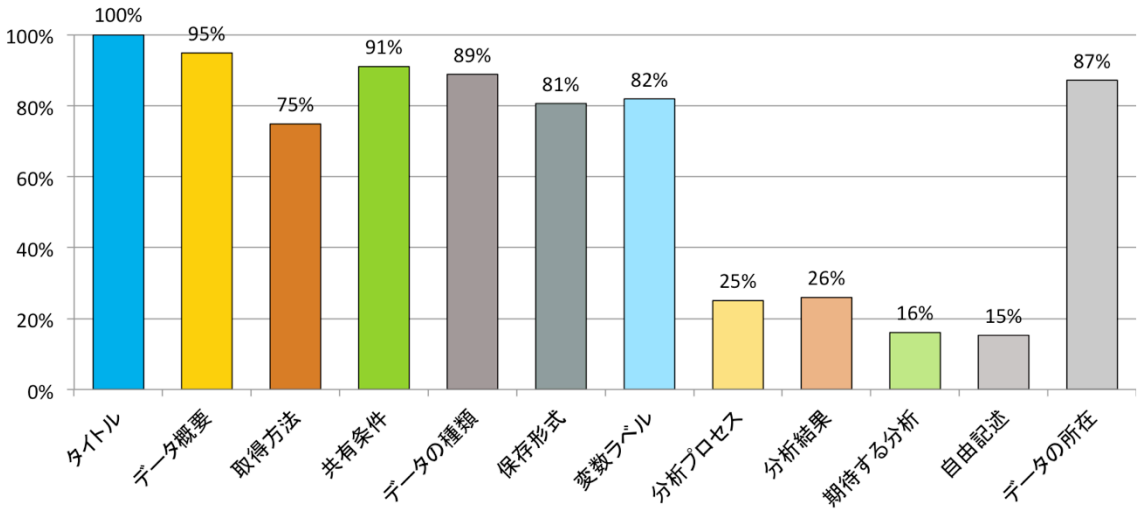

図 5-14　登録 DJ の記入率（本実験で用いる 799 件を対象）

5.4.3.2 変数ラベル推定のモデル

本節では，データの概要情報からデータに含まれる可能性のある変数ラベルを推定するためのモデル化について議論する．モデル化によって得られる機能は，変数ラベル（VL）が未知であるデータ概要（OD）から，そのデータに含まれる可能性のある変数ラベルを推定し，変数ラベルの集合を得ることである．以下に，本研究の基本的なアイデアとモデルの詳細を述べる．

- データ概要の類似度の利用（モデル 1）：OD の類似度が高ければ，データが保有する VL も類似しているという仮定に基づいたモデルである．VL が未知である OD を検索クエリとし，学習データに含まれる DJ の OD と比較して類似度の高い OD と含まれる VL をスコアリングする．
- 変数ラベルの共起性の利用（モデル 2）：VL の共起性を考慮したモデルである．「緯度」と「経度」という VL は位置情報関連の DJ に同時に登場する頻度が高いという

例のように，VL には同じ DJ 内に同時に登場する頻度が高い集合が存在しているという特徴を用いている．

実装にあたり，まず，モデル 1 について詳細を議論する．例えば，「東京都が提供する街路灯に関する設置データ（OD_x）」という VL が未知である OD が与えられているとする．ここで，「岡山の行政機関が設置した街路灯の位置情報（OD_1）」及び「東京都における年齢別の人口割合（OD_2）」という OD を持つ 2 つの DJ が存在するとする．このとき，OD_1とOD_2がOD_xと似ている度合いを類似度と定義し，それぞれ$similarity(OD_x, OD_1)$，$similarity(OD_x, OD_2)$と書くこととする．それぞれの OD を形態素解析によって単語に分けるとすると，OD_xとOD_1は「街路灯」と「設置」という共通の索引語を有しているため，OD_2と比較して類似度が高いと言える（$similarity(OD_x, OD_1) > similarity(OD_x, OD_2)$）．モデル 1 に従うと，VL が未知である$OD_x$は$OD_1$が保有する VL と同じ VL，すなわち，「緯度」，「経度」，「光束」，「管理番号」，「行政機関名」，「地域名」を有している可能性が高いと言える．つまり，モデル 1 により VL が未知の OD から，含まれている可能性の高い VL の集合を取得できる（図 5-15）．

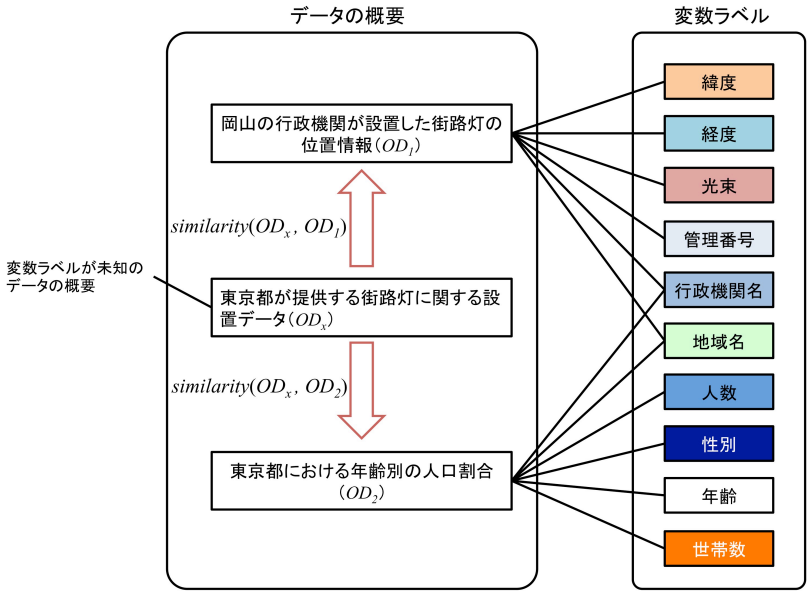

図 5-15　データ概要（OD）の類似度を考慮したモデルの例（モデル 1）

続いてモデル 2 について議論する．モデル 2 は上述したように，VL の共起性を考慮したモデルとなる．例えば，「東京都内にある企業の構成員データ（OD_3）」に含まれる変数ラベルは「年齢」，「性別」，「人数」，「企業名」であるとする（図 5-16）．一方で，図 5-15 で例示した「東京都における年齢別人口の割合（OD_2）」に含まれる変数ラベルは「行政機関名」，

「地域名」,「人数」,「性別」,「年齢」,「世帯数」であった．すると，これらの 2 つの DJ は，「人数」,「性別」,「年齢」の VL を共通して持っていることになる．つまり，「人数」,「性別」,「年齢」の VL は同時に取得される可能性が高い VL の集合であるということができ，すなわち共起性が高い VL 集合と言える．同様に，「岡山の行政機関が設置した街路灯の位置情報（OD_1）」及び「東京都における年齢別の人口割合（OD_2）」の DJ では，「行政機関名」と「地域名」という 2 つの VL を共通して有しているため，「行政機関名」と「地域名」は同じデータ内で同時に登場する頻度が高い，すなわち共起度が高い VL の集合であると言うことができる．さらに，OD_2とOD_3はどちらもある領域における人口という人の属性を内包したデータの概要であり，OD_1とOD_2は行政機関が収集しているデータという特徴を有したデータの概要であることを考えると，同時に登場する頻度が高い VL の集合を有する OD は類似している類似していると言える．

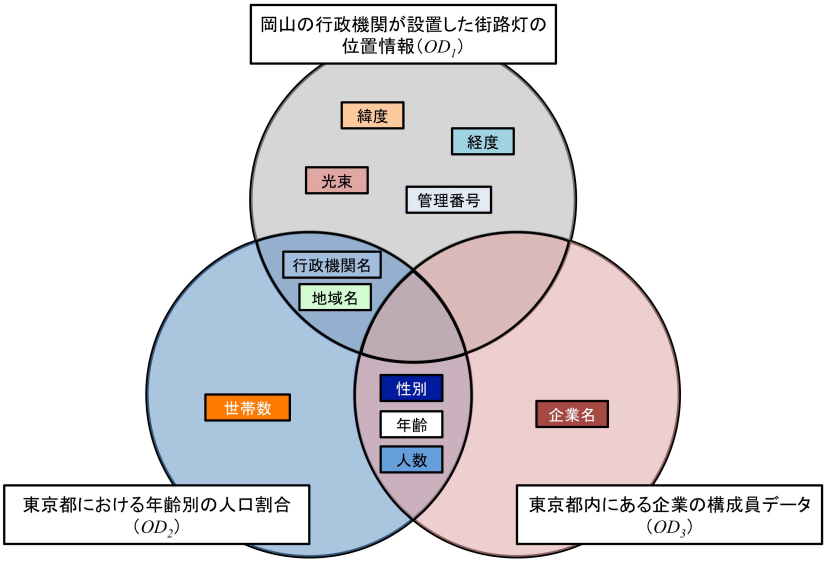

図 5-16　変数ラベル（VL）の共起性を考慮したモデルの例（モデル 2）

VL が未知である OD から，VL を推定するだけであれば，モデル 1 のみで十分と考えられるかもしれないが，推定性能がデータ概要の記述量に大きく影響されてしまう可能性がある．つまり，単語数の少ない OD をベクトル空間モデルを用いて表現したとき，記述量の多い OD とのみ類似度が高くなり，それらに含まれる特定の VL 集合のみが推定結果に現れてしまう恐れがある．また，本研究の前提にあるように，データ保有者がデータ市場に公開したい情報のみが DJ に記述されるため，登録された DJ には必ずしも全ての VL が含まれているとは限らない．そして，データには取得者の癖や取得意図が大きく影響するため，

取得可能でも分析に用いない VL はデータとして取得されず, 欠損している可能性も考えられる．さらに，類似度の高い内容を表すデータであっても OD に含まれる単語の表記揺れにより，類似度が低くなってしまうことが懸念される．例えば,「年齢」,「性別」,「身長」,「体重」の VL を有した「アメリカの子どもに関するデータ」と「日本における児童の統計情報」の DJ の OD には共通の単語が含まれておらず，データ自体の類似度が高いにも関わらず, OD 内に共通する単語を含んでいないために，類似度の値が低く見積もられてしまう．類似度を計算するためには類語関係や同義関係を考慮したシソーラスなどの高度な辞書が必要となる．

以上の問題に対して，モデル 2 に示した VL の共起性を用いることで，ある DJ に欠けている VL を他の DJ の VL を用いて補完することができ, VL 未知の OD から含まれている可能性のある VL 集合を取得するのに有効に作用するものと考えることができる．

5.4.4 変数クエスト（VARIABLE QUEST）の設計

本節では前節の議論を踏まえてモデル 1 及びモデル 2 を仮定し，OD の類似度及び VL の共起性から，VL 未知の OD に含まれる可能性の高い VL を推定する方法を説明し，その機能を有するアプリケーションを変数クエスト（VARIABLE QUEST）と名付け，実装する．本研究では, OD を bag-of-words モデル及びベクトル空間モデル（Salton et al., 1975; Turney & Pantel, 2010）を用いる. なお, 前処理として, 形態素解析器 MeCab（Kudo & Matsumoto, 2000）を用い，OD を分かち書きしたのち，助詞，助動詞，句読点などの記号類は不用語として除外した．

5.4.4.1 実装

まず，モデル 1 に基づき，OD の類似度を計算する単語-OD 行列Mを作成する．OD 集合に出現する単語の集合に対して任意の単語の OD における出現頻度をベクトル空間モデルに基づいて表現する．具体的には，OD 内のW件の単語を行ベクトル，D件の OD を列ベクトルとして，単語-OD 行列M（$|W|\times|D|$）とする（式(5.4.4.1.1)）．OD の列ベクトル$\boldsymbol{od}_j$の要素v_{ij}は, 単語t_iがデータ概要OD_jに出現する回数から重み付けした値を表す（式(5.4.4.1.2)）．なお，行列及びベクトルの右上の添字Tは転置を表し，ベクトルは太字で表現する．

$$M = (\boldsymbol{od}_1, \cdots, \boldsymbol{od}_j, \cdots, \boldsymbol{od}_D) \tag{5.4.4.1.1}$$

$$\boldsymbol{od}_j = (v_{1j} \ \cdots \ v_{ij} \ \cdots \ v_{Wj})^{\mathrm{T}} \tag{5.4.4.1.2}$$

続いて，VL の集合に対して，OD における VL の出現回数（0 または 1）をベクトル空間に表現する．図 5-15 で表したように，DJ には OD とそれに含まれている VL が紐付いている. OD に紐付くV件の変数ラベルを行ベクトルとし, D件の OD を列ベクトルとした VL-OD 行列R（$|V|\times|D|$）を作成する（式(5.4.4.1.3)）．OD の特徴ベクトル$\boldsymbol{od'}_j$の要素r_{ij}は，変数ラ

ベルvl_iがデータ概要OD_jに出現する回数を表す（式(5.4.4.1.4)）．

$$R = (\boldsymbol{od}'_1, \cdots, \boldsymbol{od}'_j, \cdots, \boldsymbol{od}'_D) \tag{5.4.4.1.3}$$

$$\boldsymbol{od}'_j = (r_{1j} \ \cdots \ r_{ij} \ \cdots \ r_{Vj})^{\mathrm{T}} \tag{5.4.4.1.4}$$

以上の手順により求めた単語-OD 行列M（$|W|\times|D|$）及び VL-OD 行列R（$|V|\times|D|$）から，単語-VL 行列E（$= MR^{\mathrm{T}}$）（$|W|\times|V|$）を作成する．これのプロセスは，i番目（$1 \le i \le |V|$）のD次元の OD 空間にある VL の特徴ベクトルを行列Mによって，W次元の単語空間に写像することに相当する．変換後の単語-VL 行列Eの要素e_{ij}は式(5.4.4.1.7)で表され，i番目の単語t_iがk番目のデータ概要od_kに含まれる頻度（v_{ik}）と，j番目の変数ラベルvl_jがk番目のデータ概要od_kに含まれる頻度（r_{kj}）の積の総和である．すなわち，単語t_iと変数ラベルvl_jの両方を含む OD の件数を表している．また，この単語-VL 行列Eは，単語・OD・VL を頂点とする 3 部グラフ構造における隣接行列の積に相当し，行列Eの成分e_{ij}は単語t_iの頂点から OD の頂点を経由し，到達可能な VL の頂点vl_jに至る経路数を表している（図 5-17）．

$$E = MR^{\mathrm{T}} = (\boldsymbol{vl}_1, \cdots, \boldsymbol{vl}_j, \cdots, \boldsymbol{vl}_V) \tag{5.4.4.1.5}$$

$$\boldsymbol{vl}_j = (e_{1j} \ \cdots \ e_{ij} \ \cdots \ e_{Wj})^{\mathrm{T}} \tag{5.4.4.1.6}$$

$$e_{ij} = \sum_{k=1}^{|D|} v_{ik} \, r_{kj} \tag{5.4.4.1.7}$$

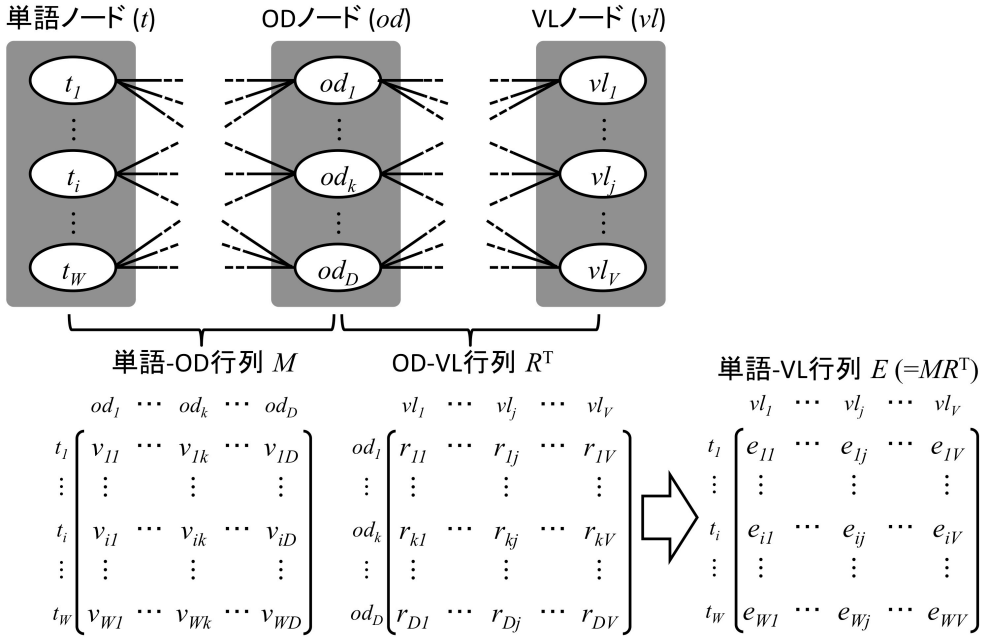

図 5-17　3 部グラフからなる単語-VL 行列E

以上のプロセスで求められた Term-VL 行列Eによって，モデル 1 が構築された．これにより，VL が未知の DJ の OD（OD_x）が与えられたときの OD を形態素解析による前処理を

施し，OD のW次元の特徴ベクトル$\boldsymbol{od}_x$を取得し，単語-VL 行列Eで表される VL の各特徴ベクトル$\boldsymbol{vl}_j$ $(1 \leq j \leq |V|)$と類似度を比較することで，類似度の高い VL のランク付けされた集合を取得することができる．

続いて，モデル 1 に対して，VL の共起性を考慮したモデル 2 を組み合わせる．まず，同じ DJ 内に含まれる任意の 2 つの VL は 1 回共起しているとし，モデル 1 で作成した単語-VL 行列Eに合わせるため，VL 共起行列C（$= RR^{\mathrm{T}}$）（$(|V|\times|V|)$）を定義する．VL 共起行列Cの要素c_{ij}はVLの対(vl_i, vl_j)のod_s $(1 \leq s \leq |D|)$における共起度$co(vl_i, vl_j)$を表す（式(5.4.4.1.8)及び式(5.4.4.1.9)）．すなわち，式(5.4.4.1.8)及び式(5.4.4.1.9)は変数ラベルvl_iと変数ラベルvl_jを含む DJ の数を表している．なお，$|vl_i|_{od_s}$はod_sにおけるvl_iの出現回数である．

$$c_{ij} = \sum_{k=1}^{|D|} r_{ik} r_{kj} \tag{5.4.4.1.8}$$

$$c_{ij} = co(vl_i, vl_j) = \sum_{s=1}^{|D|} |vl_i|_{od_s} \left|vl_j\right|_{od_s} \tag{5.4.4.1.9}$$

単語-VL 行列Eに VL 共起行列Cを右からかけることで，VL の共起性を考慮した単語-VL 行列EC（$(|W|\times|V|)$）を得る．単語-VL 行列ECとEの違いは，モデル 2 にある VL の共起性を考慮しているかどうかということであり，行列の成分が異なっている．行列ECの成分g_{ij}は式(5.4.4.1.10)で表され，ある単語t_iが与えられたときの OD の類似度（Eによる作用）と VL の共起性（Cによる作用）を考慮して得られる変数ラベルvl_jの値を表している．つまり，g_{ij}は，単語-第 1OD-第 1VL-第 2OD-第 2VL を頂点集合とする 5 部グラフにおいて，単語ノードに含まれる単語t_iから第 1OD ノードを経由して到達可能な第 1VL ノードに至り（行列Eによる作用），続いてそれらの第 1VL ノードから第 2OD ノードを経由して第 2VL ノードのvl_jに至る（Cによる作用）経路数と同義である（図 5-18）．

$$g_{ij} = \sum_{m=1}^{|V|} \left(\sum_{k=1}^{|D|} v_{ik}\, r_{km} \right) \left(\sum_{l=1}^{|D|} r_{ml}\, r_{lj} \right) \tag{5.4.4.1.10}$$

以上により，VL が未知の DJ の OD（OD_x）が与えられたときの OD を形態素解析による前処理を施し，OD のW次元の特徴ベクトル$\boldsymbol{od}_x$を取得し，単語-VL 行列ECで表される VL の各特徴ベクトル$\boldsymbol{vl}_j$ $(1 \leq j \leq |V|)$と類似度を比較することで，類似度の高い VL のランク付けされた集合を取得することができる．実際には，OD_xを特徴ベクトル$\boldsymbol{od}_x$（W次元）に変換し，単語-VL 行列ECに含まれる VL の特徴ベクトル$\boldsymbol{vl}_j$との類似度（$similarity(\boldsymbol{dj}_x, \boldsymbol{vl}_j)$）を算出し，VL 未知の OD（$OD_x$）に対して類似度の高い VL の集合を取得する仕組みとなる．

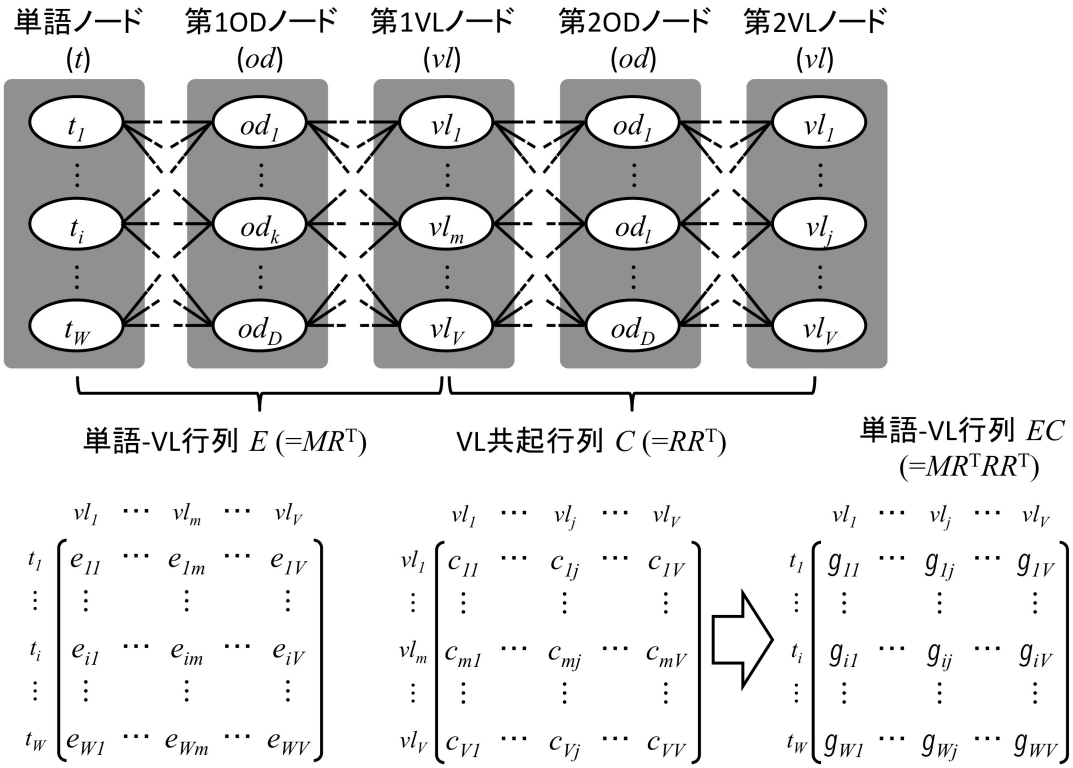

図 5-18　5 部グラフからなる単語-VL 行列EC

5.4.4.2 訓練データ

本研究では訓練データとして, 2016 年 3 月までに集められた 799 件の DJ を用いる（表 5-11）．それぞれの DJ は 1 つの OD といくつかの VL で構成されている．VL の種類は総計 3,216 種類である．すべての OD に含まれるテキストをコーパスとして辞書を作成する．コーパス及び辞書の作成には gensim[27]を用いる．句読点や記号類を不要語とし，名詞，形容詞，副詞，動詞を形態素解析によって原形に直し，2 回以上出現するもののみ抽出した．OD のコーパスは 1,935 種類の単語で構成されている．1 件の DJ は平均して 5.34 件の変数ラベルを持つ．形態素解析器には MeCab[28]を用いた．データ概要内の特徴的な単語の重み付けには tf-idf（term frequency – inverse document frequency）（Salton & Buckley, 1988）を用いる．tf-idf を用いることによって，OD に含まれる特徴的な語の重みを増やし，全体的に頻度の高い単語の重みを下げる．なお，次節で説明する本研究の実験でも同様のデータを訓練データとして用いる．

[27] http://radimrehurek.com/gensim/
[28] http://taku910.github.io/mecab/

表 5-11　訓練データ（コーパス）

データジャケット数	799
データ概要（OD）に含まれる平均単語数	39.5
データジャケットに含まれる平均変数ラベル（VL）数	5.34
OD に含まれる総単語数	30,767
OD に含まれる単語の種類	1,935
データジャケットに含まれる総変数ラベル数	4,160
VL の種類	3,216

5.4.4.3 具体例

作成した単語-VL 行列E及びECの出力結果の詳細について具体例を用いて説明する. 例として以下に示す 3 つの OD をクエリとして，OD の類似度を考慮したモデルである単語-VL 行列Eを用いた結果，そして，VL 共起性を考慮した単語-VL 行列ECの出力結果上位 10 件を表 5-12，表 5-13，表 5-14 に示す．なお，訓練データにはこの OD は含まれていない.

1. 日本を訪れた外国人観光客が飲食店で飲んだビールの消費量及び金額についてのデータ．外国人の国籍や年齢，入店人数などが含まれている.
2. 日本のある地域における年ごとの人口の推移を表すデータ.
3. 東京都内のバスの運行状況とバス停に関するデータ．バスの到着時刻などの情報が含まれている.

出力例 1 では，行列E及びECにてほぼ同等の VL が出力されており，推定結果は入力 OD に含まれる可能性が高いと考えられるものが出力された．行列Eの結果では「出身国」という VL が最も高い類似度となっているが，これは「鯖江市の外国人住民数」という DJ の OD との類似度の高さから表出したものである．また，出力例 2 では，類似度の値が上位の VL は行列E及びECでほぼ同等であるが，行列Eの結果では「農家総人口」など，農業従事者における VL が出力されていることから, 類似度の高い訓練データの OD が影響しているものと考えられる．一方，VL 共起性を考慮した行列ECでは，他の DJ に含まれる VL の同時に登場する性質により, OD に含まれる可能性の高い VL 集合が得られている. 出力例 3 では，行列E及びECにてほぼ同等の VL が出力されているが, 出力結果において類似度が高い上位の VL と比較して, 類似度の低い VL は OD に対して関連性が低くなっていると考えられる.

以上の 3 つの例から，OD の類似度を考慮したモデル 1 による単語-VL 行列Eを用いることによって，VL が未知の OD から含まれる可能性の高い VL の集合を得ることができることが分かった．また，OD の類似度だけでなく，VL の共起性を考慮することによっても，

データに含まれる可能性の高い VL を推定することができる可能性があることが分かった.

表 5-12　VL 集合の出力例 1

単語-VL 行列Eの利用		単語-VL 行列ECの利用	
VL	類似度	VL	類似度
出身国	0.370624	アクティビティの満足度	0.279646
アクティビティの満足度	0.279646	訪日旅行者の属性（年代）	0.279646
訪日旅行者の属性（年代）	0.279646	消費額	0.279646
消費額	0.279646	再訪意向	0.279646
再訪意向	0.279646	在日中に滞在した場所	0.279646
在日中に滞在した場所	0.279646	アクティビティの経験有無	0.279646
アクティビティの経験有無	0.279646	旅程を通した満足度	0.279646
旅程を通した満足度	0.279646	在日中に滞在した期間	0.279646
在日中に滞在した期間	0.279646	消費対象	0.279646
消費対象	0.279646	訪日旅行者の属性（性別）	0.279646

表 5-13　VL 集合の出力例 2

単語-VL 行列Eの利用		単語-VL 行列ECの利用	
VL	類似度	VL	類似度
人口（女）	0.285944	年齢区分（5 歳毎）	0.286694
人口（男）	0.285944	人口（女）	0.282354
人口（総数）	0.285944	人口（男）	0.282354
年（5 年毎）	0.268982	人口（総数）	0.282354
前回増減	0.268982	出生数	0.280215
流入人口	0.268982	死亡数	0.280215
農家総人口	0.256225	転入者数	0.268426
農業就業人口	0.256225	死亡者数	0.268426
農業就業人口（男）	0.254874	転出者数	0.268426
農業就業人口（女）	0.254874	人口	0.266166

表 5-14　VL 集合の出力例 3

単語-VL 行列Eの利用		単語-VL 行列ECの利用	
VL	類似度	VL	類似度
バス停名	0.330722	区間（乗車～降車）	0.211213
便名	0.330722	月日	0.211213
区間（乗車～降車）	0.234942	時刻	0.205443
月日	0.234942	消費電力	0.196205
時刻	0.179549	バス停名	0.188274
年	0.151929	便名	0.188274
防災ハンドブック	0.149460	金額	0.181002
災害時の対処法	0.149460	高度	0.177940
経度	0.148784	店名	0.173953
緯度	0.147792	ランプ種別	0.173820

5.4.5 実験

本節では，前節で構築したモデルを導入したアプリケーション「変数クエスト（VARIABLE QUEST)」の性能を評価する．テストデータを用いて，推定性能を比較することで評価を行う．

テストデータとして，静岡県庁が提供するオープンデータ「ふじのくにオープンデータカタログ」[29]のデータを DJ 化したもの 50 件を用いる（表 5-15）．また，テストデータには形態素解析器 MeCab を用い，行列E及びECと同様の前処理を行い，コーパスに含まれない単語が入力された場合は，それを除外する．また，テストデータに含まれる VL のうち，コーパス内の VL 集合に正解 VL が含まれない VL は除外する．

表 5-15　テストデータに用いる DJ の概要

データジャケット数	50
データ概要（OD）に含まれる平均単語数	36.7
データジャケットに含まれる平均変数ラベル（VL）数	4.70

[29] https://open-data.pref.shizuoka.jp/

以上の 50 件のテストデータの DJ の OD のみを抽出し，VL 未知の OD のデータセットを作成する．これらの VL が未知である OD をクエリとし，訓練データから作成した単語-VL 行列E, ECを用いて VL の各特徴ベクトルから返される類似度付き VL 集合を比較することで手法の評価を行う．類似度の計算にはコサイン類似度（$sim(\boldsymbol{od}_x, \boldsymbol{vl}_j) = \boldsymbol{od}_x \cdot \boldsymbol{vl}_j / |\boldsymbol{od}_x||\boldsymbol{vl}_j|$）を用いる．

また，比較手法として，OD に含まれる単語をクエリとして，単語の一致によって VL の集合を取得する方法（本研究では The String Matching（TSM）と呼ぶ）を用いる．例えば，前節で例示した「日本を訪れた外国人観光客が飲食店で飲んだビールの消費量及び金額についてのデータ．外国人の国籍や年齢，入店人数などが含まれている．」という OD（OD_x）であれば，OD を形態素解析によって分かち書きし，bag-of-words として取得する（$OD_x = \{t_i | i \in \mathrm{N}\}$）．そして，OD に含まれる単語$t_i$に一致するコーパス内の重複を含んだ VL 集合（$\mathcal{V} = \{vl_j | 1 \leq j \leq |V'|\}$）（$|V'|$は重複を含んだ VL の数である）から取得する．すなわち，「国籍」，「年齢」，「入店人数」，「消費量」，「金額」を VL として取得する．また，取得された VL のうち, 取得回数が多い順にスコアを付けて, ランク付きの結果を得る. なお, TSM では形態素解析器 MeCab を用い，行列E及びECを得るのと同様の前処理を行う．

本実験では, Information Retrieval 分野の検索集合評価でよく使われる Precision, Recall 及び F measure を用いる．適合率 Precision（P）及び再現率 Recall（R）は式(5.4.5.1)と式(5.4.5.2)として定義し，推定結果として返される上位 10 件が各クエリ OD に含まれるものであるか評価する．なお，各式の TP は true positives，FP は false positives，FN は false negatives を表す．また，P と R の調和平均 F measure は式(5.4.5.3)と定義する．50 件のテストデータの各クエリに対する F measure を算出し，単語-VL 行列E，EC及び TSM の値を比較する．

$$P = TP/(TP + FP) \tag{5.4.5.1}$$

$$R = TP/(TP + FN) \tag{5.4.5.2}$$

$$F = 2PR/(P + R) \tag{5.4.5.3}$$

続いて，行列E及びECの評価を行うために，Average Similarity（AS）を定義する．ランク付き推定結果を評価する方法に Mean Average Precision（MAP）があるが，これは主に検索結果の順位を評価する方法である（Buckley & Voorhees, 2000）．本研究が対象にする DJ では，各 OD に紐づく VL 同士に順位は存在しない．例えば，天候データの中の「緯度」，「経度」，「天候」の VL は互いに順位なく平等に含まれている．そのため，本実験では推定結果の順位ではなく, OD への類似度を評価する AS を式(5.4.5.4)と定義する. 式中のV_{od_q}はod_qに含まれる正解 VL 集合を意味する. また, $rel(od_q, vl_p)$は, vl_pが

正解 VL 集合に含まれているとき（$vl_p \in V_{od_q}$）1 であり，そうでないときは 0 となる関連度を表す指標である．この評価方法についても，最終的に 50 件のテストデータの各クエリに対する AS の平均を取ることで Mean Average Similarity（MAS）を算出し，行列E及びECを比較する．

$$AS_{od_q} = \frac{1}{\left|V_{od_q}\right|}\sum_{p=1}^{|V|}\left(sim\left(\boldsymbol{od_x}, \boldsymbol{vl_p}\right) \cdot rel(od_q, vl_p)\right) \tag{5.4.5.4}$$

5.4.6 結果と考察

5.4.6.1 F measure の比較

テストデータから作成した各クエリに対して出力された上位 10 件の中で正解した VL の数から Precision，Recall そして F measure を算出し，行列E，EC及び TSM を比較した（表 5-16）．その結果，行列E，ECの出力結果の F measure は TSM と比較し，高い性能があることが分かった．特に，行列Eの性能は TSM の 2.14 倍であり，行列ECでは 1.78 倍となった．なお，算出した F measure を比較したところ，単語-VL 行列E及びECにおいて対応のある t 検定を用いたところ，有意な差は見られなかった（$t(98) = -1.23, p = 0.110$）．Precision 及び Recall についても同様に，単語-VL 行列E及びECの推定結果において有意な差は見られなかった（Precision: $t(98) = 0.937, p = 0.351$, Recall: $t(98) = 0.989, p = 0.325$）．つまり，推定結果の上位に含まれる VL 集合の評価は単語-VL 行列E及びECにおいてほぼ同程度であるということができる．

表 5-16　推定結果の評価（平均値±標準偏差）

	F measure	Precision	Recall
行列E	0.235±0.178	0.174±0.131	0.401±0.331
行列EC	0.196±0.183	0.146±0.133	0.332±0.337
TSM	0.110±0.091	0.082±0.068	0.185±0.173

5.4.6.2 Average Similarity の比較

続いて，行列E及びECの評価するために，類似度が計算された VL 集合の MAS を比較する．MAS は VL 集合がクエリである OD にどの程度近いかを評価する指標である．つまり，行列E及びECの MAS を比較することで，VL の共起性を考慮したモデル（モデル 2）がどの程度 OD に対する VL の類似度に影響しているかを調べることができる．

F measure の比較と同様に 50 件のテストデータに対して MAS を比較したところ, 単語-VL 行列E及びECにおいて有意な差が見られた（$t(98) = 9.52, p < 0.01$）（表 5-17）．F measure の比較で分かるように，推定結果上位の VL 集合の評価はほぼ同じであったにも関わらず, AS を比較すると推定された VL の AS は行列ECを用いた結果のほうが行列Eと比べて向上していることが分かる．テストデータの件数では，50 件中 48 件において，行列ECを用いて推定された正解 VL 集合の AS が向上した．

表 5-17　AS の評価（平均値±標準偏差）

	AS	$\overline{AS}$
行列E	0.329±0.113	0.069±0.014
行列EC	0.399±0.095	0.111±0.016
p-value	**	**

**: p<0.01, *: p<0.05, $n.s.$: 有意差なし

しかし, 行列ECによって, 非正解 VL 集合の類似度も同時に向上している可能性がある．そこで，非正解 VL 集合の AS を$\overline{AS}$として式(5.4.6.2.1)を定義し，比較に用いる．式中の$\left|V \cap \overline{V_{od_q}}\right|$は VL 集合から$od_q$に含まれる正解 VL 集合を除いた集合の要素数を意味する．つまりクエリに対して関係のない VL の個数を表している．また，$unrel(od_q, vl_p)$は，vl_pが正解 VL 集合に含まれていないとき（$vl_p \notin V_{od_q}$）1 であり，そうでないときは 0 となる非関連度を表す指標である．

$$\overline{AS_{od_q}} = \frac{1}{\left|V \cap \overline{V_{od_q}}\right|} \sum_{p=1}^{|V|} \left(sim(\boldsymbol{od}_x, \boldsymbol{vl}_p) \cdot unrel(od_q, vl_p) \right) \qquad (5.4.6.2.1)$$

以上の式から求める$\overline{AS}$をテストデータ 50 件に適用し，行列EC及びEの非正解 VL 集合の AS（$\overline{AS}$）を比較した．結果，表 5-17 に示すように，クエリod_qに対して行列Eと比較して，行列ECの非正解 VL 集合の類似度の向上が見られた（$t(98) = 9.52, p < 0.01$）．この結果は，VL の共起性を考慮したモデルの導入により，正解 VL 集合だけでなく，非正解 VL 集合の OD への類似度も上がったということを示している．

そこで, 各クエリにおける行列ECによる類似度から行列Eによる類似度を引いた値を向上値と定義し，類似度のみを考慮した行列Eから，VL の共起性を考慮した行列ECによる作用によって正解 VL 集合の類似度（AS）と非正解 VL 集合の類似度（$\overline{AS}$）の向上度合いを比

較した．AS と$\overline{AS}$の各クエリに対する向上値を対応のある t 検定を用いて比較した結果，AS の向上値が$\overline{AS}$と比較して有意に高いという結果が得られた（AS 向上値 : 0.071, $\overline{AS}$向上値 : 0.042, $t(98) = 4.47, p < 0.01$）．本結果は，行列ECによる作用は行列Eによる作用と比較して，非正解 VL 集合の類似度も同時に向上させているが，正解 VL 集合の類似度の方が改善される度合いが大きいことを意味する．すなわち，VL の共起性を考慮したモデル（モデル 2）によって，正解 VL 集合の OD に対する類似度が向上することが分かった．

5.4.6.3 考察

実験の結果から，OD には必ずしも VL を表す単語が含まれているとは限らないことが示唆される．すなわち，欲しいデータに関する概要情報から，単語の一致のみで変数について検索し，必要な変数を発見することは困難であると言うことができる．そこで，本実験の結果より，データ概要が似ているデータは同じ変数ラベルを有していると仮定するモデル 1，そして，変数ラベルの共起性を仮定したモデル 2 を用いることによって推定性能を向上させることができることが分かった．

しかし，行列E及びECの性能について，F measure の結果としては低い値であるように思えるかもしれない．しかし，本研究が扱っている変数ラベルはコーパス内に 3,216 種類存在している．さらに，コーパス内の VL の総数は 4,160 件であり，単純計算でも 1 種類の VL 当たり約 1.28 件のみ含まれていることになる．すなわち，ほとんどの VL の出現回数は 1 回程度である．ほとんど 1 回しか登場しない VL をたかだか 40 単語程度で構成されている OD から推定可能としているところに本研究の成果であると言うことができる．

また，本研究では OD における類似度を考慮したが，VL においては考慮していない．なぜなら，VL は「緯度」や「経度」のように 1 語程度で構成されているため，ベクトル空間モデルではほとんどの特徴量が 0 となる問題があるからである．しかし，同じ意味を表す VL でも，「場所」と「所在地」，「名前」と「名称」などのように異なる書き方で記述されているものも存在する今後の研究では，VL の意味や類語関係を考慮したモデルを導入する必要があるだろう．

本研究では，データの概要説明であるデータ概要の類似度に加え，変数ラベルの共起性に注目し，推定手法の向上に役立てた．文書内の単語の共起性や論文の共著関係など，様々な共起性が存在する．本研究は，データの変数ラベルにも互いに登場する頻度が高いものが存在するということを考え，データジャケットとして取得されたデータに含まれる変数ラベルの共起性を議論した．先行研究では，データジャケットを用いたデータ利活用方法検討支援にデータジャケットの概要情報及び変数ラベルのキーワードの一致性のみに着目している（Ohsawa et al., 2013 など）が，単語の一致のみでは変数に関する情報の不足によ

り，本来結合する可能性のある DJ 同士が未結合となってしまうという問題が存在する．本研究により，変数ラベルが欠けていても，データ概要を用いた推定手法を用いることでデータ同士の潜在的な結合可能性を論じることが可能となると考えられる．

自然言語のフリークエリ（free text queries）から変数ラベルを含む詳細な情報の検索が可能となる．先行研究では，自然言語によるクエリからデータ概要内のキーワードを手がかり語として関連する DJ を検索する仕組みは第 4 章 4.4 節の DJ ストアとして提案されていたが，変数ラベルは検索対象に含まれていなかった．本手法によって，新たなデータ取得者が，過去の事例からどのような変数ラベルを用いることが適当であるか知ることができる．また，2016 年 11 月現在公開されている DJ は 1,100 以上件存在し，変数ラベルの種類は約 5,000 種類である．そのため，新たにデータを取得したい人が，取得したいデータについての情報を自然言語で検索することで，過去の DJ から取得すべき変数ラベルを知ることができる．つまり，人間の認知負荷を軽減することができると考えらえる．

5.5 本章のまとめ

本章では，データ利活用に関わる諸要素が意思決定者のデータ利活用方法検討のプロセスにどのように現れるのかを観察し，データから価値を策定するプロセス及び価値を発見され得るデータの特徴について論じた．従来研究では，データ利活用に関わるステークホルダーやリソースなどの要素について検討する意思決定者のプロセスにおいて，どのような要素がどのような過程を経て決定されるのかについて明らかとなっていなかった．

そこで，5.1 節では，データ利活用のための実行動シナリオ生成支援手法アクション・プランニング（AP）について説明し，5.2 節にて，データ利活用に関わるステークホルダーやリソースなどの要素について考察する意思決定者のプロセスにおいて，どのような要素がどのような過程を経て検討されるのかについて実験的に考察した．実験により，シナリオ生成プロセスにおいて，要素の追加や削除という矛盾を解消する行動が観察され，シナリオ生成プロセスには仮説推論における非単調性が現れることが分かった．また，実験的データ市場においてデータに文脈を付与するプロセスでは，データの組合せだけでなく，ステークホルダーやリソースといったデータに関わる諸要素の関連性を考慮した検討が重要であるという示唆が得られた．

そして，5.3 節では，第 4 章 4.3 節及び 4.4 節にて検討したデータ利活用知識の構造化を拡張し，人間の意思決定を支援するシナリオ創出手法 AP によって生成されたシナリオの構造化と再利用する仕組みににについて議論した．そして，データ利活用による新規事業創出のためのシナリオ生成支援手法として，ステークホルダー表出と関係推定システム Resource Finder（リソースファインダー）を実装し，文脈によって異なるステークホルダーのシナリオへの関係を推定可能であることを実験的に評価した．

さらに，5.4 節では新たにデータを取得する意思決定者の支援手法として変数ラベル推定方法について議論し，VARIABLE QUEST（変数クエスト）を開発し，その性能を評価した．従来手法では，新たにデータを取得したい人がどのような変数を取得することが，意思決定に役立つのかという情報は蓄積されてこなかった．もし，データ取得後に取るべきであった変数が明らかとなったとき，新たにコストをかけて取得しなければならないという問題が生じ得る．本手法により，新たにデータを取得し，意思決定に役立てたいと考えるデータ取得者に対し，過去に蓄積された DJ の情報を用いることで，どのような変数を取得することが意思決定において重要であるのかという知見を示すことができる．

本章では，データ利活用シナリオの構造化による行動計画立案支援手法について論じた．次章では，データ利活用シナリオを元に実際にデータの入手，分析，分析結果に基づく行動を行った際のギャップ及び実行動の過程について実験的に考察する．

第６章　実装的データ市場におけるデータ利活用プロセス

第 4 章，第 5 章では，実験的データ市場をワークショップ形式で実施することで実データ市場の仕組みの理解と提案手法の評価を行った．本章では，実験的データ市場の枠組みを拡張し，データの入手，分析，課題発見とフィードバックという実社会とのインタラクションを含んだ実装的データ市場を観察する二つの応用実験を行う．

第 5 章 5.1 節の実験により，AP のシナリオ生成プロセスにおいて，矛盾を解消する行動が起こることが筆記行動の分析により分かった．すなわち，IMDJ で検討されたデータの組合せ案から AP において実行動シナリオを生成するというデータ利活用方法検討においては，矛盾が発生する可能性が示唆された．本章では，シナリオ生成から実際のデータ分析の間のギャップについて考察を行う．6.1 節では，被験者の分析シナリオとそれを元にした実際のデータの分析のプロセスを観察し，定量的・定性的な評価を行う．また，6.2 節ではデータ利活用方法の検討のあと，シナリオを元に分析を行い，実際のステークホルダーとのインタラクションからデータの入手とフィードバックを得るプロセスについて観察した事例について，考察する．そして，6.3 節では，本章のまとめを行う．

6.1 応用実験１：分析シナリオに基づく実分析

6.1.1 分析シナリオと実分析のギャップ

分野を横断したデータ交換及び利活用に対する期待が高まっているものの，データの入手・利活用などのあらゆる実行動には多大なコストが生じる．高度なセンサーの利用やデータのプライバシー及びセキュリティの観点から，良質なデータの価値は高くなり，価格は高騰する可能性がある．また，複雑なデータであれば複数の分析手法を検討した上で，時間をかけて分析結果を解釈することが必要である．さらに，分析で得られた結果を元に新事業を創出するときには，ステークホルダーや必要なリソースの検討が必要になる．つまり，意思決定の判断材料となる情報の入手，分析などのすべての行動にはコストがかかる．そこで，事前にどのようなデータを入手し，どのような分析によってどのような仮説が検証できるのかというシナリオを立て，評価することが重要であると考えられる．

しかし，データ分析シナリオとシナリオを元にして実際に分析をして結果を得ることには隔たりが存在する．例えば，分析時に必要なデータが入手不可能であることが発覚するなど，矛盾が生じた際に代替手段で目的を達成しなければならない．また，計画時には矛盾がなくとも，コストをかけて実際に手に入れたデータが想定したものではなく，計画を練り直すなどの手戻りが発生することが考えられる．人間の製品設計プロセスにおいて手戻りは重大なリスクとして認識されている．開発初期段階でコストや技術リスクを考慮しないような不適切な設計プロセスによって，生産における矛盾や衝突が生じる可能性があ

ることが指摘されている（Koga & Aoyama, 2004; 古賀・青山, 2010 など）．データ利活用においても，実際のデータ分析が計画通りに進まないことによる手戻りの発生やコストの問題は避けて通ることができないと考えられる．

本節では，応用実験として，実際のデータ利活用の現場を模したワークショップを開催し，データ利活用アイデアの創出，分析シナリオの生成，そして実データ分析と分析結果を得る一連のデータ利活用プロセスを観察する．特に，データ分析シナリオと実際の分析行動の間のギャップに着目し，データ利活用の現場において意思決定者を支援すべき部分について議論する．第 4 章及び第 5 章にて，IMDJ 及び AP にて述べ，これらの手法によっていくつかの新しいデータ利活用ツールの創出や意思決定に役立った事例などの成果が報告されていることを説明した．しかし，従来研究では，分析シナリオに基づいて行動し，実際のデータ分析と結果を得るプロセスの観察と検討は不十分であった．本研究の目的は，データ利活用におけるシナリオと，シナリオに基づいた実行動の間のギャップの要因を理解し，データ利活用支援のための示唆を得ることである．実験では，データ利活用において最も重要な要素の一つであるデータの分析行動に着目し，研究者及び社会人 18 人を対象としたワークショップを開催した．シナリオ生成及び実分析において参加者が困難と感じる部分やシナリオと実分析の間の矛盾が生じ得る箇所，そして矛盾解消行動について，第 5 章 5.1 節の実験で用いたデジタルペンを用いた筆記行動の分析，ワークショップ中の参加者の行動の観察，そしてアンケートによって取得する．

6.1.2 実験

ワークショップのテーマは「組織及び個人の疲労とミス検知のためのデータ分析案の創出」である．ワークショップは 3 つのフェーズによって構成されている．初めに，参加者はデータ利活用方法検討手法である IMDJ によって DJ からテーマに基づく要求及びソリューションを考案する．ここでは仮説の元となる要求と，仮説検証のためのデータ分析で用いるデータ及びツールの組み合わせ案を議論する．続いて，2 から 3 人程度のグループでアクション・プランニングを実施する．AP では，IMDJ で創出された要求をシートに従って要求を仮説化し，仮説を検証するために必要な分析プロセスを変数ラベルと擬似コード形式記述を用いて表現した分析シナリオを創出する．最後に，シナリオに基づいて実際のデータを取得し，データ分析を行うフェーズを実施する．参加者には DJ に対応する実データが与えられ，シナリオ実現に必要な分析を行う．参加者は研究者，実業家及び工学系の学生の合計 18 人が参加した．

6.1.2.1 IMDJ フェーズ

はじめに，参加者に本実験のテーマと内容を教示する．データ利活用方法検討ワークショップ IMDJ にて，既存の DJ からテーマに基づいて IMDJ は 60 分間，要求とソリューションを考案する．IMDJ の実施 10 人程度が適当であるため，2 つのテーブルで実施した．IMDJ の実施手順の概要は以下である．

1. 事前に IMDJ に用いるデータジャケットを 39 件の範囲で準備し，データジャケット間で共有された単語を結ぶグラフからマップを作成した（参加者の共通参照情報として KeyGraph（Ohsawa et al., 1998）を用いたが，組み合わせのヒントの提示が趣旨であり，可視化ツールの効果比較は論じない）．可視化した図は図 6-1 に示した．
2. 参加者は実在のデータ利用者（可能なら自身）の利害を主張する立場から，要求を提示する（約 15 分）．
3. 提案者の立場から，データジャケットを組み合わせることで要求を満たすソリューションを創出する．なお引き続き，要求を提示可能である．
4. 利用者は自身の要求を満たすソリューションが創出された場合，ゲーム中の架空通貨を提案者に支払う（ソリューションの評価として計上する）．
5. 手順 2 から 4 を約 45 分間繰り返し，最も架空通貨を得たソリューションを精緻化アクション・プランニングの精緻化対象とする．

本実験では，2 つのグループの中からそれぞれ 3 件のソリューション（合計 6 件）を選定し，AP によって分析シナリオを創出した．

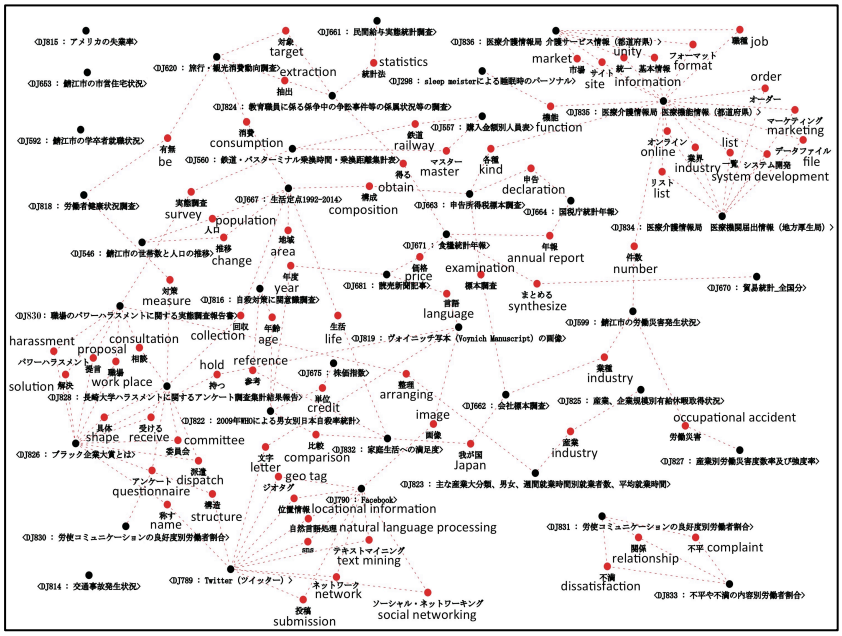

図 6-1　実験に用いたシナリオマップ

6.1.2.2 AP フェーズ

本実験の AP は，分析シナリオの生成支援を目的として設計した．AP シートは図 6-2 に示すように，要求仮説化，擬似コード記述，データ・ツール表出化の 3 部構成である．

要求の仮説化とは，「データを用いて検討可能な粒度に要求を精緻化すること」である．例えば，「2020 年開催予定の東京オリンピックは大変な混雑が予想されるため，混雑を解消したい」というのは要求であるが，データを用いて検証できる粒度であるとは言い難い．一方で，「観光地の回りに飲食店が多いと混雑少なくなる」や「国籍によって興味のある観光地が異なる」というものは，仮説と呼ぶことができる．なぜなら，前者は「飲食店の位置情報」と「観光地の入場者数データ」の組み合わせによって検証が可能であり，後者は「観光客の属性データ」と「観光客の旅行先の訪問地に関するデータ」などを組み合わせることにより検討可能であるからである．つまり，仮説とはデータによって検証可能な粒度で記述された仮定であると考えることができる．本実験では，図 6-2 に示す図を用いて，本ワークショップのテーマである「組織及び個人の疲労とミス検知のためのデータ分析案の創出」に対して実施したデータ利活用方法検討ワークショップ IMDJ にて創出された要求及びソリューションに対し，「ステークホルダー」，「状況」，「背景要因」，「概要」の 4 つの項目を記述することで仮説を導く方法を導入した．本実験において AP は分析シナリオと実分析の相違点を観察することが目的のため，AP シートの項目に関する比較は扱わない．

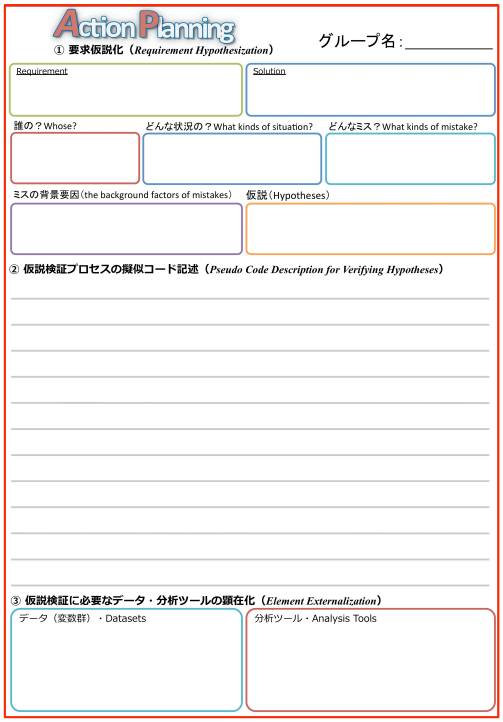

図 6-2　実験に用いた AP シート

擬似コード形式記述とは，仮説検証に必要な変数ラベルの組合せや分析ツールの組み合せを検討する論理的なプロセスを記述することである．実データがなくとも，DJ に含まれる変数ラベルの組み合わせから分析の道筋を考えることができることは第 4 章 4.1 節にて述べた．擬似コード記述では，入出力変数ラベルから期待する分析結果を得るための論理的な分析ロジックを考えることができる．例えば，「地図上に街路灯データをマッピングし，安全・安心なルートを提案するシステムを開発する」というソリューションには以下の 2 つの要求が背景にあるとする．

- 要求 1：明るい道を歩くことで安心する．
- 要求 2：暗い道は犯罪に巻き込まれる危険性がある．

また，以上の要求から導かれる仮説は，以下であるとする．

- 仮説 1：昼よりも夜間の方が犯罪が多く発生する
- 仮説 2：夜間の暗いエリアでは街灯があるエリアと比べて犯罪が多く発生している

このような仮説を検証するプロセスを擬似コード形式で記述するのが擬似コード記述である．図 6-3 は上記の仮説を検証するプロセスの擬似コード記述の例である．各行が関数を表しており，上の行で宣言した変数ラベルを入力として，変数ラベルの組合せから必要な変数ラベルが出力される手順を記述していく．擬似コード化プロセスにて表出化した変数ラベルを含む DJ の実データを取得し，検討した手順に従って分析することで期待する分析結果及び仮説の検証が可能となる．

```
1. 必要な変数の取得と定義

get 犯罪発生時間、犯罪発生場所(緯度)、犯罪発生場所(経度)
get 街灯光束、街灯設置場所(緯度)、街灯設置場所(経度)
plot 犯罪発生場所(緯度、経度) on 地図(緯度、経度)
plot 街灯設置場所(緯度、経度) on 地図(緯度、経度)
define 犯罪(夜間)= 犯罪 during (0 AM～6 AM) or (5 PM～0 AM) from 犯罪発生時間
define 犯罪(昼間)= 犯罪 during (6 AM～5 PM) from 犯罪発生時間


2.「昼よりも夜間の方が犯罪が多く発生する」という仮説の検証
compare 犯罪(夜間) with 犯罪(昼間)
analyze 相関 comparing 犯罪(時間帯、発生件数)


3.「夜間の暗いエリアでは街灯があるエリアと比べて犯罪が多く発生している」という仮説の検証
define 明るさ(6段階) from 街灯光束
define エリア明るさ(範囲) = 明るさ(6段階) × 3m(半径)
combine エリア明るさ(緯度、経度、範囲) with 犯罪(緯度、経度、夜間)
compare 明るさ(0～2)の犯罪(夜間) with 明るさ(3～5)の犯罪(夜間)
```

図 6-3　擬似コード記述の例

データ・ツールの表出化とは，データ分析に必要なデータあるいは分析ツールを記述することである．仮説化あるいは擬似コード記述プロセスで必要と気付いたデータ・ツールを記述するメモ欄としての役割を果たす.

また，本実験において AP は分析シナリオと実分析の相違点を観察することが目的のため，チーム毎の分析シナリオ生成過程における思考時間を第 5 章 5.2 節の実験で用いたデジタルペンによる筆記データを用いて，筆記推移を追跡し，チーム毎に比較する．

6.1.2.3 データ分析フェーズ

各チームに DJ に対応する実データを配布し，AP で作成した分析シナリオを元に仮説を検証する分析フェーズを 60 分間実施した．分析中に，仮説の検証に必要なデータやツールが足りないことに気付いた際には，Web 上で入手可能なデータに限り，アクセスして利用することを許可した．もし分析過程で仮説を変更しなければならないことに気付いた場合は，チーム内で再度検討し，検証すべき仮説を変更することを許可した.

分析フェーズ終了後，1 チーム 5 分程度の発表と 5 分の質疑応答を経て，もっともよいデータ分析結果を得たグループに 1 人 1 票の相互評価を行った．参加者 18 人に加え，グループに参加しなかったファシリテーター1 人を加えた 19 人が評価を行った.

6.1.3 結果と考察

データ利活用アイデアの創出，データ分析シナリオの生成，そして実データ分析と分析結果を得る一連のデータ利活用プロセスを観察した結果を説明する．IMDJ フェーズでは，グループ 1 で 18 件の要求，12 件のソリューションが創出された．一方，グループ 2 では 22 件の要求と 18 件のソリューションが創出された．IMDJ ではグループによって創出される要求やソリューションの数に大きな差はなかった．IMDJ で高評価を得て，AP でシナリオ化が検討されたソリューションは表 6-1 に示す.

その後，表のソリューション及び要求を AP において仮説化し，分析シナリオを作成したところ，6 つのチームのうち 4 つのチーム（B，C，D，F）において仮説化によって IMDJ のソリューションから変化した．例えば，チーム B の「余裕のある人に仕事を振る」というソリューションは仮説化において，要求が提起される状況や関係者の表出化から，「介護の現場におけるミスを減らすための適正人数の算出と不満分析によるミスマッチ率の計算」に変化した．一方，チーム A 及び E は IMDJ のソリューションから大きく変化していない.

表 6-1　高評価ソリューションの満たした要求及び組合せた DJ のタイトル

チーム	要求	ソリューション	組合せた DJ のタイトル
A	自殺を減らしたい	男女別自殺率のデータを利用してできるかぎり自殺を防止する	・ 2009 年 WHO による男女別日本自殺率統計
B	プロジェクト初期に実施プロセスや役割分担を明確にしてストレスを減らしたい	余裕のある人に仕事を振る	・ 労使コミュニケーションの良好度別労働者割合 ・ 家庭生活への満足度
C	本当にストレスの溜まる業界はどこか知りたい(国民性･海外)	ブラック職種・地域・健康割合・自殺	・ 労働者健康状況調査 ・ 2009 年 WHO による男女別日本自殺率統計 ・ 医療介護情報局 医療機能情報（都道府県） ・ 産業別労働災害度数率及び強度率
D	-	パワハラキーワードランキングの公表	・ 長崎大学ハラスメントに関するアンケート調査集計結果報告 ・ 職場のパワーハラスメントに関する実施調査報告書
E	-	株価と有給の相関関係を調べる	・ 株価指数 ・ アメリカの失業率
F	働きながらも低ストレスの人の生活パターンを知りたい	生活定点とストレスの関係	・ sleep meister による睡眠時のパーソナルデータ ・ 生活定点 1992-2014

続いて，AP フェーズを実施した結果を示す．AP ではデジタルペンを用いて筆記データを取得し，図 6-4 に例示したように可視化した．表 6-2 は可視化した筆記の軌跡を元に AP のシートの記入順序，データ及びツールの追加・修正，そして分析結果の評価である．

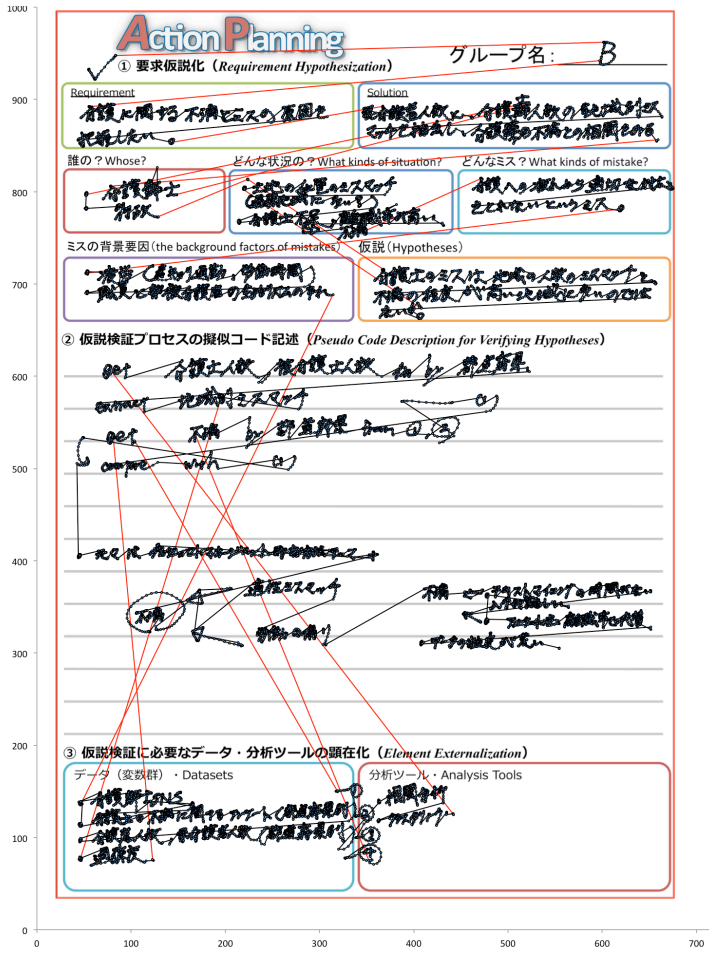

図 6-4　デジタルペンを用いた筆記軌跡の可視化例

表 6-2　各チームにおける行動と分析結果の評価

チーム	AP シート記入順序	IMDJ→AP		AP→データ分析		評価
		データ	ツール	データ	ツール	
A	1 → 2	+	+	+	+	6
B	1 → 3 → 2 → 3 → 2 → 3 → 2	+	+	−	−	0
C	1 → 2 → 3	0	+	0	−	4
D	1 → 3 → 2	−	+	+	+	1
E	1 → 2	0	+	+	0	8
F	1					0

1：要求仮説化部，2：擬似コード記述部，3 データ・ツール表出化部

+：要素の追加，-：要素の削除，0：追加・削除なし

表から分かるように，チーム B 及びチーム F の評価は他のチームと比較して低い．チーム B は筆記行動から分かるように，擬似コード記述部とデータ・ツール表出化部において頻繁に修正行動（要素の追加あるいは削除）を行っていることが分かる．つまり，分析シナリオ内の矛盾に気付き，頻繁に矛盾解消行動を行っている．第 5 章 5.2 節の実験結果を踏まえると，十分な長さの時間と矛盾解消行動のコストを認めれば，評価に値するシナリオが創出できることが分かっている．しかし，今回は 60 分という全体で統一された限られた時間の中での分析シナリオの創出であったため，チーム B はシナリオの改善に十分な時間が確保できず，分析結果においても評価が低くなってしまったと考えられる．一方，チーム F は要求仮説化部において議論が収束せず，定められた 60 分間を費やしていたため，十分に分析シナリオを練ったとは言えず，分析結果の評価が低くなってしまったものと考えられる．

IMDJ, AP そして実際の分析におけるデータ及びツールの追加･削除の行動に着目すると，全てのチームにおいて IMDJ で検討していたソリューションから AP において分析シナリオを創出する過程で分析ツールを追加していることが分かる．すなわち，IMDJ においてはデータに用いる具体的な分析ツールが未検討であったが，分析プロセスを議論する AP において分析ツールの必要性に気付いたものと考えることができる．また，興味深いのは，データの追加だけでなくデータの削除が起こっていることである．IMDJ だけではデータの組合せのみが検討されていたのが，AP のプロセスで具体的に変数ラベルの組合せから必要なデータ及びツールが選別されたと言うことができる．また，IMDJ と AP だけでなく，AP によって創出された分析シナリオに従って実分析を実施した際にもデータ及び分析ツールの追加や削除が観察された．この結果は，シナリオと実際の分析活動の間にもギャップが存在することを示している．つまり，分析シナリオに沿って実際に分析すると，期待していたデータの形式でなかったため，結論やデータ分析の方法を変更しなければならないという矛盾解消行動が現れたものであると言える．特に興味深いのは，データが手に入らないことが分かっても，代替データで当初の目的を達成しようとするグループ，データの形式が異なっていたことから，当初の目的とは異なる代替目的を設定してデータ分析を行ったグループが見られたことである．

また，データ分析シナリオを創出する部分で，擬似コードによる変数ラベルの組み合わせから期待する分析結果を得るプロセスを記述する部分が最も難しいと回答された．これの部分は，必要なデータから期待する分析結果を得るプロセスであり，データ利活用においてもっとも重要な部分であると考えられる．今後の研究では擬似コード記述について，アプリケーションなどによる支援を検討する必要があるだろう．

6.2 応用実験 2：実行動における課題発見とフィードバック

本節の内容は大澤ら（2017）の単行本「データ市場（近代科学社）」の「第 5 章：アクション・プランニング」に含まれる「5.4　シナリオの実行と評価」にて同様の内容を公表予定であるため，図表及び記述内容の一部を除外して記述する．

6.2.1 シナリオの実行と評価

シナリオは人間の意思決定を支援する行動計画であるが，シナリオとそのシナリオを元にした実行動には，ギャップが生じ得ることが前節の実験によって分かった．例えば，分析時に必要なデータが入手不可能であることが発覚した場合は，代替手段で目的を達成しなければならない，あるいは目的自体を変更する必要が発生する．また，シナリオには矛盾がなくとも，実際に手に入れたデータが想定したものではなく，シナリオを練り直さなければならなくなるなどの手戻りが発生することがあり得る．

そこで，本節では，前節の実験の結果を踏まえ，実際のステークホルダーとのインタラクションからデータの入手とフィードバックを得るプロセスを複数回繰り返すことによる効果を観察し，シナリオ実施による課題発見とフィードバックについて調査した結果について述べる．

6.2.2 実験

本実験では研究者及び社会人を対象とし，IMDJ 及び AP のワークショップを開催した．このワークショップには 28 名の参加者を集め，事前に収集した 30 件の DJ から，前節と同様の手順で IMDJ 及び AP を実施した．AP は実社会におけるステークホルダーとのインタラクションを想定し，分析に特化したシートではなく，第 5 章 5.1 節にて説明した最も基本的な AP の手順に沿って行うシートを用いた．

90 分の IMDJ から，40 件の要求と 35 件のソリューションが創出され，創出された 35 件のソリューションから，架空通貨によって特に高く評価されたソリューションを 14 件選出し，参加者に 2 人 1 組となってもらい，90 分間の AP によって戦略的シナリオを生成してもらった．そして，生成された 14 件の戦略的シナリオを，Hayashi et al.（2014）の研究の評価指標と同様に，新規性・有用性・実現性を考慮して相互評価してもらった．以上の手順によって高評価を得た戦略的シナリオは「街路灯データ」と「地図データ」を組み合わせて創出された「地図上に街路灯データを入れ，明るいルートを検索する防犯アプリケーション」となった．続いて，創出された戦略的シナリオを元に実際にデータを入手し，実社会のステークホルダーとのインタラクションから課題を発見し，フィードバックを得るプロセスを複数回繰り返す作業を行った．

6.2.3 結果と考察

6.2.3.1 ステップ 1

IMDJ のソリューション及び AP によって精緻化されたシナリオの仮説は,「街路灯の多い場所は明るい．そして，明るい場所は安全である」というものであった．すなわち，街路灯データの中で最低限必要な情報は街路灯の設置位置を表す変数ラベルである「緯度」,「経度」であった.「緯度」と「経度」が分かれば，地図上に街路灯の位置をプロットすることできるからである.

このシナリオを実現するためのステップとして，始めにオープンデータとして公開されている静岡県の道路照明灯データ（第 5 章 5.4 節の変数クエストで用いた「ふじのくにオープンデータカタログ」より取得）を利用した．本データは静岡県庁道路保全課が CSV 形式で提供しているものである．データに含まれている変数ラベルは「事務所名」,「設置位置」,「緯度」,「経度」,「X 座標」,「Y 座標」,「路線名」,「管理番号」,「照明灯種別」の 9 種類であり，各変数ラベルの値として静岡県庁が管理する県内の街路灯情報が含まれている．組み合わせる地図データとして，Google が提供している Google Maps API v3[30]を用いた．前述の 2 つのデータを組合せ，各街路灯の「緯度」と「経度」の値から地図上にプロットした．以上に加え，さらにルート検索機能を実装し，最も街路灯が多いルートを検索するアプリケーションを開発した．このアプリケーションの実装には約 2 週間を要した（表 6-3）.

本アプリケーションの評価を得るために，シナリオのステークホルダーの一つである「行政」に該当する東京都文京区役所の街路灯担当者に問い合わせ，インタビューを行った．そして，本アプリケーションによる機能を評価してもらうとともに，文京区役所が管理する街路灯データの入手交渉を行った．インタビューの結果，街路灯の維持管理は県，市，町がそれぞれ担当しており，ステップ 1 で開発した街路灯情報をプロットしただけのアプリケーションでは不十分であることが分かった．つまり，本アプリケーションの機能では，県道にある街路灯のみが可視化されており，ステークホルダーの一つである「住民」のニーズを満たすことができないという課題が明らかとなった．また，街路灯は光束によって明るさが異なっており，本アプリケーションでは，必ずしも明るいルートであるとは言えないという評価を得た.

一方，文京区が管理する街路灯のデータはオープンデータとなっておらず，文京区役所は一般に街路灯データを提供していなかったが，静岡県の街路灯データを用いたシナリオと本アプリケーションを担当者に見せたところ，研究目的の利用に限り，文京区内の街路灯データの提供を得ることができた.

[30] https://developers.google.com/maps/

表 6-3　ステップ 1 の分析及び行動に関わった要素及びその概要

ステップ	要素名	概要
ステップ 1	ステークホルダー	開発者
		評価者「行政」
	ツール	JavaScript
	データ（変数ラベル）	Google Maps API v3
		静岡県街路灯データ（緯度，経度）
	期間	2 週間

6.2.3.2 ステップ 2

ステップ 2 では，ステップ 1 で行政の街路灯管理者からのフィードバックを踏まえ，変数ラベル「光束」を含めた新たな課題について開発者内で再度シナリオの検討を行った．ステップ 1 で開発したアプリケーションを利用するユーザー（住民）の立場に立ったとき，ユーザーが地図に表示されたルート上の街路灯の本数を数え，明るいと判断したルートを決定することは現実的ではない．従って，ルートを色によって明るさを識別可能なアプリケーションとするという改良案に至った．また，さらにアプリケーションの評価者としてステークホルダーに「市民」を加え，アプリケーションの改良と実験を行った．また，ステップ 1 における交渉によって入手した文京区の街路灯データは CSV 形式であり，含まれる変数ラベルは「照明種別」,「管理番号」,「柱種」,「適合ランプ」,「町名」,「緯度」,「経度」の 7 種類であり，文京区内に存在する私道を除いた道路に設置された街路灯データが記載されていた．ステップ 1 と同様に Google Maps API v3 を用い,「適合ランプ」の種類から「光束」を数値化した値を加え,「緯度」,「経度」を用いて地図上に可視化した．

続いて，ステークホルダーである「住民」からのフィードバックを得るため，被験者 8 人を募り，GPS 機能を備えた携帯端末を持って，夜間に特定の区間を歩いてもらう実験を実施した．実験後，明るいと感じた区間，そして暗いと感じた区間を歩行した地図上にいめしてもらうというアンケートを行った．続いて，携帯端末から取得した GPS データを元に被験者が移動したルートと明るさに関する主観評価の結果と実際の光束による明るさをサポートベクターマシン（SVM）法及び判別分析法で比較したところ，判別率は 75%程度という結果を得た．さらに，被験者実験及びアンケートによるフィードバックによって，夜間は比較的明るく，大きい通りを通る傾向があることが分かった．また，本ステップが完了するまでに約 3 週間を要した（表 6-4）．なお，SVM 法及び判別分析法は Python 及び R のパッケージを用いた．

表 6-4　ステップ 2 の分析及び行動に関わった要素及びその概要

ステップ	要素名	概要
ステップ 2	ステークホルダー	データ提供者「文京区役所」
		開発者
		評価者「住民」
	ツール	JavaScript
		Python
		R
		サポートベクターマシン
		判別分析法
	データ（変数ラベル）	Google Maps API v3
		GPS による位置情報（緯度，経度，時間）
		文京区街路灯データ（緯度，経度，光束）
	期間	3 週間

6.2.3.3 ステップ 3

ステップ 2 の実験による被験者からのフィードバックから，大通りの明るさが重要であることが分かった．大通りは区役所ではなく，都庁が管理しているというフィードバックは文京区役所から得ていた．そこでステップ 3 では，より精緻な分析結果を得るために都道に関するデータの取得のインタビューを行った．ステップ 1 と同様に，東京都庁建設局に街路灯データの利活用方法と分析結果を見せ，本アプリケーションによる機能を評価してもらうとともに，東京都が管理する街路灯データの入手交渉を行った．交渉の結果，一般的に共有不可能である都道の街路灯に関するデータの提供を得た．このデータには，文京区の街路灯データと同様に，「緯度」，「経度」，「光束」の情報が含まれていたため，文京区の街路灯データとの粒度を揃える加工を施し，再度 SVM 法及び判別分析法を被験者のデータに適用した．

その結果，判別率は約 82%に向上した．つまり，新たなデータの追加により，より高い精度で明るい道を識別することができたのである．さらに，ステップ 2 に加え，データとステークホルダーが増加したと同時に，以上のプロセスを完了するまでにおよそ 5 週間を要した（表 6-5）．

表 6-5　ステップ 3 の分析及び行動に関わった要素及びその概要

ステップ	要素名	概要
ステップ 3	ステークホルダー	データ提供者「文京区役所」
		データ提供者「東京都庁」
		開発者
		評価者「住民」
	ツール	JavaScript
		Python
		R
		サポートベクターマシン
		判別分析法
	データ（変数ラベル）	Google Maps API v3
		GPS による位置情報（緯度，経度，時間）
		文京区街路灯データ（緯度，経度，光束）
		東京都街路灯データ（緯度，経度，光束）
	期間	5 週間

6.2.3.4 シナリオ実施における課題発見とフィードバック

データ利活用方法検討ワークショップ IMDJ によって要求を満たすデータの組み合わせから創出されたソリューションは，AP によってステークホルダーや必要なリソースを考慮した戦略的シナリオとなる．このシナリオ生成過程において，矛盾解消行動が行われることは第 5 章 5.2 節の実験によって分かっていたが，実際の分析における課題発見とフィードバックについては議論が十分ではなかった．そこで本節では，実際のステークホルダーとのインタラクションからデータの入手とフィードバックを得るプロセスを 3 回繰り返すことによる効果を観察し，考察を行った．

各ステップにおけるステークホルダーからのフィードバックにより，新たなステークホルダーの存在が明らかとなったり，必要なデータの存在や検討すべき新たな課題について気付きが得られることが分かった．図 6-5 は，全 3 回のステップの期間と各要素の増加を図示したものである．図からわかるように，ステップを経るに連れて関連する要素が増加するとともに，期間が伸びていることが分かる．また，ステップを経ることでステークホルダーや必要なデータ，分析ツールなどが増大し，目的達成までの期間は延長されるものの，精度の高い分析結果を得られる可能性は高くなることが分かった．本節の実験では，3 回の

ステップのみを追跡して調査したが，この実験以降のステップを経ることでも，さらに要素が追加され，目的達成までの期間が延長されるとともにアプリケーションの精度も向上するものと考えらえる．なぜなら，3 回目のステップの成果を研究者及び役所の職員に見せて議論したところ，夜間の明るさはコンビニなどの夜間に営業している店舗の影響も大きいことが分かった．さらに，本実験に用いたデータには私道における街路灯の情報は含まれていない．さらに，区内では街路樹が障害となり，街路灯の明るさが定められた値以下となっているエリアがあるというフィードバックも得られている．また，人間が夜間に選択する道は，明るさだけでなく道の幅，高低などの変数ラベルも影響を及ぼす可能性が考えられる．これらのデータを当該機関に交渉，あるいは実地調査によって入手することによってさらに分析結果の精緻化とアプリケーションの機能向上が期待できるだろう．

また，文京区役所及び東京都庁では，当該データの活用方法を示したシナリオを提示することで，一般には共有できないデータを範囲を限定して共有可能となるという重要な示唆を得た．この現象は，創出されたデータ利活用案（ソリューション）を見たデータ保有者は，該当データを共有するための有益な情報（所有者や入手方法など）を新たに提供する傾向が見られたという先行研究（大澤, 2014）の報告と合致するものであり，IMDJ 及び AP が有効に作用したものと考えることができる．

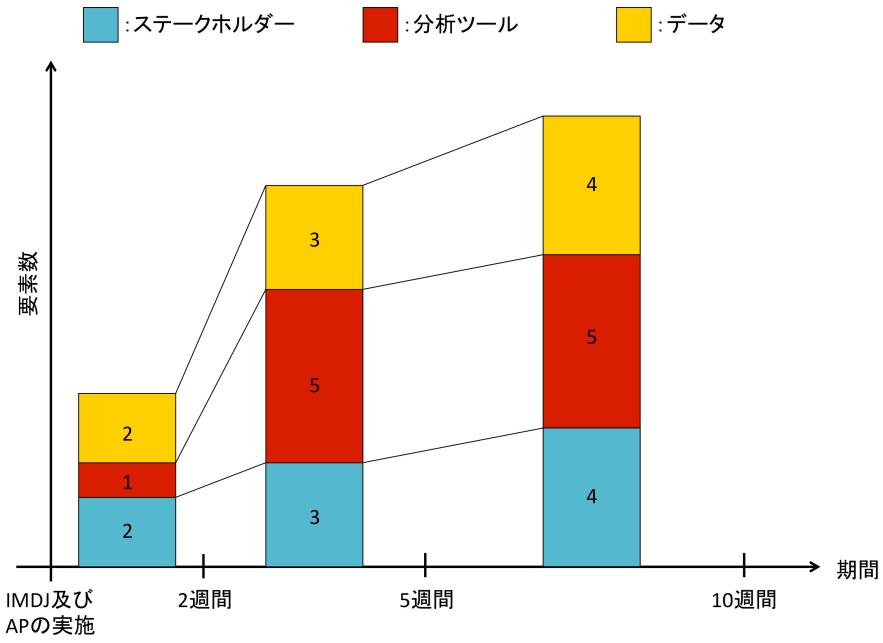

図 6-5　全 3 回のステップにおける要素と期間の増加

6.3 本章のまとめ

データ利活用方法検討ワークショップ IMDJ によって要求を満たすデータの組み合わせから創出されたソリューションは，AP によってステークホルダー，必要なリソース，変数ラベルを考慮し，分析シナリオ及び戦略的シナリオとなる．このシナリオ生成過程において，要素の追加あるいは削除といった矛盾解消行動が行われることによってシナリオとして精緻化されることは第 5 章の実験によって分かっていた．しかし，第 5 章の研究は実験的データ市場における参加者の行動を観察することによって得られた知見であった．つまり，シナリオを元に実際のデータ分析と結果を得るプロセスについては検討が十分ではなかった．そこで，本章では，第 3 章にて定義した実験的データ市場の枠組みを拡大し，実際のデータの入手，実データの分析，課題発見とフィードバックという実社会とのインタラクションを含んだ実装的データ市場におけるプロセスを観察した．6.1 節，6.2 節の応用実験によって，実装的データ市場においては，実際にシナリオを元に行動すると，途中で仮説を修正したり，期待する結果が得られないといった理想と現実のギャップに遭遇することが明らかとなった．

6.1 節の実験では，実際のデータ利活用の現場を模したワークショップを開催し，データ利活用アイデアの創出，データ分析シナリオの生成，そして実データ分析と分析結果を得る一連のデータ利活用プロセスを観察した．特に，IMDJ のデータ利活用案と AP による分析シナリオの間だけでなく，分析シナリオと実際の分析行動の間にもギャップが存在することが分かった．つまり，データ分析シナリオを創出したが，分析シナリオに沿って実際に分析すると，期待していたデータではなかったり，期待する分析結果が得られなかったなどにより，結論やデータ分析の方法を変更するという行動が見られた．特に興味深いのは，目的が達成できない状態に陥ったときに，2 つのパターンの行動を被験者が取ったことである．一つ目は，データが手に入らないことが分かっても，代替データで当初の目的を達成しようとする行動であり，二つ目は，期待していたデータとは異なっていたことから，当初の目的とは異なる代替目的を設定してデータ分析を行うという行動である．現実における問題解決過程は，不完全な知識を仮定して推論を進めていくことで問題解決に至るプロセスである．そのため，矛盾が生じた場合，今までの仮定を棄却・修正するという行動が現れることが知られている．データ市場における実分析も同様であり，実際の分析は様々な試行錯誤によって結果を得るプロセスであり，分析計画と実データ分析の間にはギャップが存在する．しかし，上流設計部にあたるデータ入手以前の分析シナリオのプランニングを十分に議論していれば，比較的評価の高い分析結果が得られる可能性が高まることが実験によって分かった．また，データ分析シナリオを創出する部分で，擬似コードによる変数ラベルの組み合わせから期待する分析結果を得るプロセスを記述する部分が最も難し

いと回答された．これの部分は，必要なデータから期待する分析結果を得るプロセスであり，データ利活用においてもっとも重要な部分であると考えられる．今後の研究では，変数ラベルの組合せから結論を導くプロセスにおいて，支援手法を検討する必要があるだろう．

6.2 節の応用実験では，IMDJ にて創出された「地図上に街路灯データを入れ，明るいルートを検索する防犯アプリケーション」を AP によってシナリオとして精緻化し，実際のステークホルダーとのインタラクションからデータの入手とフィードバックを得るプロセスを複数回繰り返すことによる効果について観察した．6.1 節では一度の分析に留まっていたが，6.2 節の実験ではデータの分析とフィードバックにサイクルを 3 回実施した．実験により，シナリオと実際の分析行動の間にもギャップが存在することが分かっただけでなく，分析結果を元にした行動によって関係するステークホルダーから新たな課題の発見とフィードバックが得られることが分かった．さらに，新たなステークホルダーの存在が明らかとなったり，必要なデータの存在について気付きが得られることが確認された．すなわち，データ市場においては，シナリオに基づく行動，分析，そして評価（フィードバック）のサイクルを回すことで，目的達成のための新たな課題が発見できる他，新たなデータ，データの変数，そしてステークホルダーが追加されることが分かった．また，ステップを経ることでステークホルダーや必要なデータ，分析ツールなどが増大し，目的達成までの期間は延長されるものの，精度の高い分析結果を得られる可能性は高くなるという示唆が得られた．また，一般に共有できないデータでも，当該データの活用方法を示したシナリオを提示すれば，範囲を限定してデータの共有が可能となるという重要な示唆を得た．以上のことから，行動しながら新しい知識を取り込み，シナリオを修正しながら分析結果を精緻化するプロセスを経ることが，データ市場における新たなステークホルダーの存在を掘り起こし，新たなデータの利用方法の発見を促すことができると考えられる．

第 7 章　結論

本章では，結論として第 2 章から第 6 章の概要と研究成果，そして本研究の将来の展望についてまとめる．

7.1 各章の概要

7.1.1 第 2 章の概要

第 2 章では，データの人間社会における役割と位置づけについて述べ，データ市場がどのような社会的要請から発展してきたのかを関連研究及び事例を示しながら明らかにした．分野を横断したデータの交換によって，既存のサービスの付加価値向上や新製品の開発などのデータ駆動型イノベーションに対する期待が高まってきているものの，様々な社会的障壁がある．このような状況において，データの公開ではなく，市場における交換という戦略によってデータ駆動型イノベーションを推進しようとする様々な形態のデータ市場が萌芽し始めた．

しかし，Web を新たなプラットフォームとするデータ市場では，データの表層的な情報を Web 上に陳列するだけのサービスに留まっており，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不十分であった．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．また，異分野のデータ結合による知識発見が期待されているものの，実際は異なる分野のどのようなデータが結合可能であり，どのような仮説を検証できるのかという知識が確立されていない．そこで，データの公開・共有を強制するのではなく，市場の原理で利用者が必要なデータを選び，所有者と交渉の末に入手することができる，イノベーション創出環境におけるデータの価値化プラットフォームであるデータ市場が提案された．以上のようなデータの価値化とイノベーション創出が有効に働く環境としてのデータ市場を整備するための支援手法の必要性について論じた．

7.1.2 第 3 章の概要

第 2 章にて論じた本研究の背景及び課題を踏まえ，第 3 章では，本研究の着眼点とアプローチについて論じ，本研究の提案手法の概要について説明した．

データ市場は今まで世の中に出てこなかったデータ，ステークホルダー，知識が登場する新しい市場であるが，昨今のデータ市場では，データの価値を策定するための背景知識が適切な形で蓄積されてこなかった．データの蓄積方法は積極的に議論されてきたが，既存の知識やモデルでは扱えないデータを含むデータ利活用知識の蓄積方法及びその評価方

法については十分に議論されてこなかったと言える．

第 3 章では，データ市場と実験的データ市場について，本研究のアプローチ方法，データの利用価値の観測方法について論じ，本研究の提案手法についての概論を述べた．

7.1.3 第 4 章の概要

第 4 章では，データ市場におけるデータ駆動型イノベーションに貢献することを目的とし，データ利活用知識構造化と検索システムによる人間のデータ利活用シナリオ生成支援手法の提案，そして実験的データ市場における参加者のプロセスの観察を行った．

4.1 節及び 4.2 節では，本研究にて用いる基礎技術として，データの概要情報であるデータジャケット（DJ），データ利活用方法検討ワークショップ Innovators Marketplace on Data Jackets（IMDJ）について概説した．4.3 節では，ユーザーのデータ利活用方法検討を支援するためのデータ記述モデルとデータジャケットの構造化について検討した．そして，4.4 節では，4.3 節の議論を踏まえ，データ利活用知識の構造化と検索システムについて論じた．データジャケットだけでなく，過去の IMDJ において議論された要求，データ利活用案によって価値を認められたデータの関係をデータ利活用知識としてモデル化し，Data Jacket Store（DJ ストア）を実装した．データ利活用知識構造化により，ユーザーが自分と異なる視点を持つ過去のユーザーが考案したデータの使い道を発見したり，過去の別の人が考案したデータ結合案に注目することによって役に立つデータジャケットを探し出すことが可能であることが評価実験により示した．

さらに，4.5 節のデータの共有条件に着目したデータの価値化の実験では，IMDJ 及び DJ ストアを用い，一般的に公開不可能な秘匿データに対する利用者の期待の高さを確認した．つまり，オープン化できないデータほど，提案者及び利用者にとって問題解決及び新ビジネス創出において有用性が認められる可能性が高いことを意味する．さらに，データに関する情報を検索するユーザーの検索行動においても，一般に共有不可能なデータの閲覧数の方が高いという傾向が見られ，共有が困難なデータの方がユーザーの興味・関心の度合いも高くなる可能性があることが分かった．

以上，第 4 章にて議論したように，膨大なデータから必要な知識を発見することが困難であり，データ市場において複数の領域にまたがって存在するデータ，ステークホルダー，ツールなどすべての要素を考慮することは難しい．それ故，意思決定者の異なる価値観や関心を持つ多様な背景知識，意図に対応して適切に設計し，構造化された知識ベースとそれを検索するシステムが必要となる．第 4 章では，データ市場における「データ」，「ソリューション」，「要求」の 3 つの要素に着目し，構造化と検索システムについて議論し，過去に検討された知識及びシナリオの再利用は新たな知識獲得に有用であることを示した．

そして，秘匿データのほうが共有可能データと比較して利用期待度が高くなる傾向から，データ市場はオープンデータに代表される公開可能データのみに閉じられた場ではなく，公開が難しい個人や企業のデータ及びその保有者を巻き込むイノベーションの場として機能し得ることが分かった．

7.1.4 第 5 章の概要

第 4 章では，データ市場における支援技術についてデータ利活用知識の構造化について論じたが，第 5 章では，データ利活用に関わる諸要素が意思決定者のデータ利活用方法検討のプロセスにどのように現れるのかという，シナリオ生成プロセスとシナリオの構造化に焦点を当てて議論した．従来研究では，データ利活用に関わるステークホルダーやリソースなどの要素について検討する意思決定者のプロセスにおいて，どのような要素がどのような過程を経て決定されるのかということについては明らかとなっていなかった．データは適切な文脈が与えられなければ人間の意思決定に役立つ情報あるいは知識にならないため，データの入手以前に利用価値を策定することは困難である．また，ビジネス機会の損失のリスクなどから，利用方法が定まっていない状態でのデータ交換は成立しにくい．そこで，IMDJ 及び AP によってデータ利活用シナリオを作成することにより，データ自体ではなく，データの概要情報からデータの活用方法を議論することができるようになる．

5.1 節では，データ利活用シナリオ創出手法アクション・プランニング（AP）について概説し，5.2 節では，データに文脈が付与されるプロセスとして，データ利活用に関わるステークホルダーやリソースなどの要素について考察する意思決定者のプロセスを観察した．事業計画立案時の筆記行動に着目した実験により，AP を用いたシナリオ生成プロセスにおいて，要素の追加や削除という矛盾を解消する行動が観察され，シナリオ生成プロセスには仮説推論における非単調性が現れることが分かった．また，実験的データ市場においてデータに文脈を付与するプロセスでは，データの組合せだけでなく，ステークホルダーやリソースといったデータに関わる諸要素の関連性を考慮した検討プロセスが重要であるという示唆が得られた．

5.3 節では，第 4 章のデータ利活用知識の構造化を拡張し，人間の意思決定を支援するシナリオ創出手法 AP によって生成されたシナリオの構造化と再利用する仕組みについて議論した．そして，データ利活用による新規事業創出のためのシナリオ生成支援手法として，ステークホルダー表出と関係推定システム Resource Finder（RF）を実装し，文脈によって異なるステークホルダーのシナリオへの関係を推定可能であることを実験的に評価した．RF により，過去に検討されたデータ利活用シナリオの再利用が，文脈によって異なるステークホルダーのシナリオへの関係を推定するのに有効に作用することが分かった．

そして，5.4 節では新たにデータを取得する意思決定者の支援手法として変数ラベル推定方法について議論し，VARIABLE QUEST（変数クエスト）を開発し，その性能を評価した．実験では，変数ラベルの類似度と変数ラベルの共起性を考慮することで，変数ラベルが未知のデータ概要からそのデータに含まれる可能性の高い変数ラベルを推定できることを実験により示した．本手法により，新たにデータを取得し，意思決定に役立てたいと考えるデータ取得者に対し，どのような変数の組合せをデータとして取得することが意思決定において重要であるのかという知見を示すことができると考えらえる．

第 5 章の議論によって，実験的データ市場において，データに文脈を与え，実社会において有用である事業とするには，データの組合せだけでなく，ステークホルダーやリソースといったデータに関わる諸要素の関連性を考慮した行動計画の検討が重要であることが分かった．そして，第 4 章の「データ」，「ソリューション」，「要求」に加え，「ステークホルダー」，「変数ラベル」に着目し，推定システムについて議論し，過去に検討されたシナリオ及びデータ概要情報の再利用は新たな知識獲得に有用であることを示した．

7.1.5 第 6 章の概要

第 6 章では，第 3 章にて定義した実験的データ市場の枠組みを拡張し，実際のデータの入手，実データの分析，課題発見とフィードバックという実社会とのインタラクションを含んだ実装的データ市場におけるプロセスを観察した．6.1 節，6.2 節の応用実験によって，実装的データ市場においても，実際にシナリオを元に行動すると，途中で仮説を修正したり，期待する結果が得られないといったギャップに遭遇し得ることが明らかとなった．

6.1 節の実験では，実際のデータ利活用の現場を模したワークショップを開催し，データ利活用アイデアの創出，データ分析シナリオの生成，そして実データ分析と分析結果を得る一連のデータ利活用プロセスを観察した．実際のデータ分析は様々な試行錯誤によって結果を得るプロセスであり，実装的データ市場における分析計画と実データ分析の間にもギャップが存在することが明らかとなった．しかし，上流設計部にあたる分析シナリオのプランニングを十分に議論していれば，比較的評価の高い分析結果が得られる可能性があることが実験により分かった．

6.2 節の応用実験では，IMDJ にて創出された「地図上に街路灯データを入れ，明るいルートを検索する防犯アプリケーション」を AP によってシナリオとして精緻化し，実際のステークホルダーとのインタラクションからデータの入手とフィードバックを得るプロセスを複数回繰り返すことによる効果について観察した．6.1 節では一度の分析に留まっていたが，6.2 節の実験ではデータの分析とフィードバックにサイクルを 3 回実施した．実験により，シナリオと実際の分析行動の間にもギャップが存在することが分かっただけでなく，

分析結果を元にした行動によって関係するステークホルダーから新たな課題の発見とフィードバックが得られることが分かった．また，一般に共有できないデータでも，当該データの活用方法を示したシナリオを提示すれば，範囲を限定してデータの共有が可能となるという重要な示唆を得た．以上のことから，行動しながら新しい知識を取り込み，シナリオを修正しながら分析結果を精緻化するプロセスを経ることが，データ市場における新たなステークホルダーの存在を掘り起こし，新たなデータの利用方法の発見を促すものと考えられる．

7.2 本研究の成果

本研究では，データ市場におけるデータ駆動型イノベーションへの貢献を目的とし，データ利活用知識構造化と検索システムによるデータ利活用支援手法の提案，そして IMDJ 及び AP という実験的・実装的データ市場におけるステークホルダーの挙動の観察から以下に示す示唆を得た．

実験的データ市場において DJ の利用とは，データ交換・売買以前のデータ利活用に対する期待の高さを表しており，経済財としてのデータの価値を保有者に提示することを意味する．つまり，データ利用者の要求を満たすデータ利活用案が分析者から提案されることにより，データの価値が定まり，データの需要が生じる．その需要によって供給の必要性が生じ，取引のための条件（価格）が調整されるという市場の原理が働く．すなわち，データ市場において，データの概要情報の共有によって秘匿データの有用性評価が可能となり，ステークホルダー間のコミュニケーションとデータの価値化を促進させると考えられる．また，利用方法などの利用価値が定まれば，データの交換・売買などの共有可能性が高まることが実験により確認された．

そして，過去に検討されたデータ利活用知識及びシナリオの再利用は新たな知識獲得と利活用方法の検討に有用であることが分かった．特に，実験的データ市場におけるデータ利活用案検討プロセスでは，既存の知識や情報だけでは解決できない問題に直面した際に，検索による新しいデータに関する情報の発見が解を促進させ得る．つまり，データに関する情報の陳列だけではなく，データ利活用知識の構造化によって発見できるデータが解発見を促し，新たな需要を喚起する可能性がある．

膨大なデータから必要な知識を発見することが困難であるように，データ市場において複数の領域にまたがって存在するデータ，ステークホルダー，ツールなどすべての要素を考慮することは難しい．それ故，意思決定者の異なる価値観や関心を持つ多様な背景知識，意図に対応して適切に設計し，構造化された知識ベースとそれを検索するシステムが必要となる．ステークホルダー及び変数ラベルの推定手法では，構造化したデータ利活用知識を再利用することにより，データ利活用において潜在的なビジネスパートナーや取得すべき変数についての情報をユーザーに提示できる可能性が示唆された．また，実験的データ市場におけるシナリオ生成過程では，データ利活用における多様な意図や制約を考慮しなければならず，アイデアを精緻にするプロセスは非単調な推論過程を経ることが分かった．また，データに文脈を付与するプロセスでは，データの組合せだけでなく，ステークホルダーやリソースといったデータに関わる諸要素の関連性を考慮した検討プロセスが重要であるという示唆が得られた．

さらに，データの共有条件に着目した実験では，一般的に公開不可能なデータに対する

利用者の期待の高さを確認した．すなわち，データ市場はオープンデータに代表される公開可能データのみに閉じられた場ではなく，公開が難しい個人や企業のデータ及びその保有者を巻き込むイノベーションの場として機能し得る．

最後に，実験的データ市場の枠組みを拡張し，実際のデータの入手，実データの分析，課題発見とフィードバックという実社会とのインタラクションを含んだ実装的データ市場におけるプロセスを観察したところ，分析プロセスの詳細な検討が実分析に影響する可能性が示唆された．また，当該データの活用方法を示したシナリオを提示すれば，範囲を限定してデータの共有が可能であるという示唆を得た．

本研究の提案手法により，データ市場において既存の知識やモデルでは扱えないデータの利活用法及び潜在的なステークホルダー，変数ラベルの発見が支援されるとともに，データ利活用知識ベースが更新されることが分かった．また，それらの手法を利用して新たに知識を獲得する意思決定者の行動が改善されることが示された．

本研究の提案手法及び得られたデータ市場のモデルによって，データに関する情報及び知識の蓄積が可能となったことで，従来の IMDJ 及び AP をデータ市場創出支援技術として大きく改良したということができる．また，本研究の提案手法によって，事業者の新しいデータの発見や異なる事業者とのインタラクションによって新規事業創出が促進されたことが報告されており，データ市場の創出と発展に貢献したものと考えることができる．

7.3 本研究の展望

本研究の長期的な目的である，データ市場におけるデータ駆動型イノベーションへの貢献のための今後の展望として，以下の 4 点が挙げられる．

1. 自然言語処理技術の応用とデータ利活用知識の拡張

   本研究では，データ利活用知識の構造化と再利用の有用性検証に重点を置いていたため，DJ ストアや Resource Finder においては単語の重み付けによるランキングは扱わなかった．同様に理由で，VARIABLE QUEST においても，bag-of-words モデルを用いたコサイン類似度による単純なモデルによる比較に留まった．ソリューション創出回数やコーパスを用いた単語の重要度を考慮するなどの自然言語処理技術によるシステムの拡張が今後の課題である．また，本研究の Resource Finder はステークホルダーに着目したが，データ利活用シナリオにはステークホルダーだけでなく，データ，分析ツール，コストなど他の様々な要素が含まれている．これらの要素についても，情報を検索するユーザーに推薦する仕組みを構築できると考えられる．また，本研究では，データ利活用シナリオに関わる諸要素のシナリオへの関係を文脈に基づいて表出し，推薦する仕組みを提案したが，検索結果に含まれる要素の相互作用や矛盾については扱っていなかった．今後の研究課題として，検索結果に含まれる要素の関係を考慮する必要があるだろう．

2. 分析事例の概要情報抽出とデータ利活用事例の蓄積

   本研究では，データ概要情報として DJ を構造化し，データ市場における様々な背景知識を持つステークホルダーを参加者として，データの組合せからデータ利活用案を検討してもらい，データ利活用知識の知識ベースを作成した．また，シナリオについても実行動及び分析に関わる要素を構造化し，知識ベースとした．しかし，本研究の知識の構造化では複雑化を避けるため，要素間の相互作用や時間順序などの情報は敢えて含めていなかった．さらに，分析の事例には得られた分析結果だけでなく，様々な試行錯誤によって結果を得るプロセスを含んでいる．今後，データ市場における知識の再利用を議論する際には，このような知識を記述するフレームワークを考えていく必要があるだろう．先行研究として，森ら（2007 など）は統計解析の学習支援のために，データの解析を記述した事例（解析ストーリー）とデータ指向統計解析環境 DoSS$^{@}d$[31]というシステムを提案している．解析ストーリーは自然

[31] http://mo161.soci.ous.ac.jp/@d/indexj.html

言語による説明文とグラフなどの図によって記述されており，XML ファイルで提供されている．だが，森らの研究は学習に主眼をおいており，すでに確立されたデータセットと分析手法の組合せに議論が留まっているため，データ市場における新しいデータや分析手法から得られる新しい知見に関しては本研究の知識ベースを援用するなどの方法が考えられるだろう．

3. DJ ストアをプラットフォームとしたデータ交換促進

データ市場では，様々な意図と背景知識を持ったステークホルダーが参加者となり，その相互作用の中でデータの価値を評価し，その評価に合う条件を設定してデータの取引を行うイノベーションの場である．本研究で開発した DJ ストアは自身の興味関心のあるデータに関する情報を検索するシステムであるが，IMDJ の参加者だけでなく，オンラインのユーザー同士のコミュニケーションによるデータ価値化の促進の効果も無視することができない重要な要素であると考えられる．今後の研究として，DJ を通して背後にある秘匿データの利活用方法を検討可能とするだけでなく，秘匿データの保有者とオンラインでコミュニケーションし，データ交換の交渉や取引を可能とするプラットフォームとしての DJ ストアの機能拡張が必要であると考えられる．

4. データ市場におけるステークホルダーの検討

第 2 章で述べたように，実社会において様々な形態のデータ市場の創生されてきている中で，データに関わる新しい職業が出てきている．本研究では，データの保有者，利用者，分析者をデータ市場に関わるステークホルダーの最小単位として議論したが，さらにデータを収集し販売するデータブローカー，そして必要なデータを選択し提示するデータキュレーターなどのステークホルダーが現れてきている．また，データ流通において悪意のあるステークホルダーの存在も指摘されており，すでにクラッカーやハッカーによって盗まれたデータの市場が形成されていると指摘する先行研究がある（Holt & Lampke, 2010）．今後の研究では，このような職業のデータ市場における役割と他のステークホルダーとの相互作用のモデル化と評価が課題となると考えられる．

参考文献

[Acquisti & Gross, 2009] Acquisti, A., and Gross, R.: Predicting social security numbers from public data, Proceedings of the National Academy of Science, Vol.106, No.27, pp.10975–10980, 2009.

[Arndt et al., 2015] Arndt, N., Ackermann, M., Brümmer, M., and Riechert, T.: Knowledge Base Shipping to the Linked Open Data Cloud, Proceedings of the 11th International Conference on Semantic Systems, pp. 73-80, Sep, 2015.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z.: DBpedia: a Nucleus for a Web of Open Data, In Aberer et al. (Eds.): The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, pp.722-735, Nov, 2007.

[秋山ら, 2013] 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.

[Bateson, 1972] Bateson, G.: Steps to an Ecology Mind, The University of Chicago Press, 1972.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O.: The Semantic Web, Scientific American, May, 2001.

[Berners-Lee, 2006] Berners-Lee, T.: Linked Data-Design Issues, <http://www.w3.org/DesignIssues/LinkedData.html>, 最終アクセス 2016 年 6 月 27 日.

[Bhardwaj et al., 2015] Bhardwaj, A., Deshpande A., Elmore, J.A., Karger, D., Madden, S., Parameswaran, A., Subramanyam, H., Wu, E., and Zhang, R.: Collaborative Data Analytics with DataHub, Proceedings of the 41st International Conference on Very Large Data Bases, Vol.8, No.12, pp.1916-1919, 2015.

[Blei et al., 2003] Blei, M.D., Ng, Y.A., Jordan, I.M.: Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol.3, pp993-1022, 2003.

[Boisot & Canals, 2004] Boisot, M., and Canals, A.: Data, Information and Knowledge: Have We Got It Right?, Journal of Evolutionary Economics, Vol.14(1), pp.43–67, 2004.

[Bollier, 2010] Bollier, D.: The promise and peril of big data, Communications and Society Program, The Aspen Institute, Washington, DC, 2010.

[Boyd & Crawford, 2012] Boyd, D., and Crawford, K.: Critical Questions for Big Data, Information, Communication & Society, Vol.15, No.5, pp.662-679, 2012.

[Buckley & Voorhees, 2000] Buckley, C., Voorhees, E.M.: Evaluating Evaluation Measure Stability, In Proc. SIGIR, pp.33-40, 2000.

[Clark, 1997] Clark, A.: Being There: Putting Brain, Body, and World Together Again, The MIT Press, Cambridge, 1997.

[Cox & Brna, 1995] Cox, R., Brna, P: Supporting the Use of External Representations in Problem Solving, Journal of Artificial Intelligence in Education, Vol.6, No.2, pp.239-302, 1995.

[Cropley, 1967] Cropley, A.: Creativity, London: Longmans, 1967.

[David, 2003] David, P.A.: The Economic Logic of 'Open Science' and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer, The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium, National Academies Press, 2003.

[Damasio, 2000] Damasio, A.: The Feeling of What Happens: Body and Emotion in the Making of Consciousness, Mariner Books, 2000.

[Dumbill, 2012] Dumbill, E.: Data Markets Compared: a look at data market offerings from four providers, < http://radar.oreilly.com/2012/03/data-markets-survey.html>, 2012, 最終アクセス 2016 年 10 月 13 日.

[DXC, 2014a] データエクスチェンジ・コンソーシアム: <http://di-d.jp/DI_20140417.pdf>, ビッグデータの先進的な利活用の推進を目的とする「データエクスチェンジ・コンソーシアム」設立のお知らせ, 2014, 最終アクセス 2016 年 6 月 27 日.

[DXC, 2014b] データエクスチェンジ・コンソーシアム: 企業保有のビッグデータの連携チャンス発見に向けてデータエクスチェンジ・コンソーシアムが東京大学・大澤幸生教授の「Innovators Marketplace on Data Jackets」を活用, http://www.data-xc.jp/release/dxc20141031.pdf, [最終アクセス 2015 年 4 月 14 日].

[Electronics and Power, 1980] Electronics and Power: New move into data market, Electronics and Power, Vol.26, No.8, pp.619, 1980.

[Ellram & Tate, 2016] 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, 2016.

[Finke et al., 1996] Finke, R.A., Ward, T.B., and Smith, S.M.: Creative Cognition: Theory, Research, and Applications, A Bradford Book, 1996.

[Freeman, 1984] Freeman, R.: Strategic Management: A Stakeholder Approach, Pitman, 1984.

[Friedman & Sunder, 1994] Friedman, D., and Sunder, S.: Experimental Methods: A Primer for Economists, Cambridge University Press, 1994.

[Goldberg et al., 1992] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D.: Using Collaborative Filtering to Weave an Information Tapestry, Communications of the ACM, Vol.35, No.12, pp.61-70, 1992.

[橋田, 2014] 橋田浩一: 分散 PDS と集めないビッグデータ, 人工知能, Vol.29, No.6, pp.614-621, 2014.

[林ら, 2007] 林勇吾, 三輪和久, 森田純哉: 異なる視点の基づく協同問題解決に関する実験的検討, 認知科学, Vo.14, No.4, pp.604-619, 2007.

[Hayashi et al., 2006] Hayashi, Y., Miwa, K., and Morita, J.: A laboratory study on distributed problem solving by taking different perspectives, Proceedings of the 28th Annual Conference of the Cognitive Science Society, pp. 333-338, 2006.

[Hayashi & Ohsawa, 2013] Hayashi, T., Ohsawa, Y.: Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game Plus Action Planning, International Journal of Knowledge and Systems Science, Vol.4, No.3, pp.14-38, 2013.

[早矢仕・大澤, 2013] 早矢仕晃章, 大澤幸生: 制約とコミュニケーションによるアイデア精緻化メソッド : アクション・プランニング, 第 12 回情報科学技術フォーラム(FIT 2013), pp.405-408, 2013.

[Hayashi & Ohsawa, 2015a] Hayashi, T., Ohsawa, Y.: Knowledge Structuring and Reuse System Design Using RDF for Creating a Market of Data, 2nd International Conference on Signal Processing and Integrated Networks, pp.566-571, 2015.

[Hayashi & Ohsawa, 2015b] Hayashi, T., and Ohsawa, Y.: Relationship between Externalized

Knowledge and Evaluation in the Process of Creating Strategic Scenarios, Open Journal of Information Systems (OJIS), Vol.2, No.1, pp.29-40, 2015.

[Hayashi & Ohsawa, 2015c] Hayashi, T., and Ohsawa, Y.: Knowledge Structuring and Reuse System Using RDF for Supporting Scenario Generation, 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES2015), Procedia Computer Science, Vol.60, pp.1281-1288, 2015.

[Hayashi & Ohsawa, 2016a] Hayashi, T., and Ohsawa, Y.: Meta-data Generation of Analysis Tools and Connection with Structured Meta-data of Datasets, 3rd International Conference on Signal Processing and Integrated Networks, 2016.

[Hayashi & Ohsawa, 2016b] Hayashi, T., and Ohsawa, Y.: Comparison between Utility Expectation of Public and Private Data in the Market of Data, 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, 2016.

[早矢仕・大澤, 2016] 早矢仕晃章, 大澤幸生: データ利活用知識構造化と再利用による検索システム: Data Jacket Store, 人工知能, Vol.31, No.5, 2016.

[Herlocker et al., 2004] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T.: Evaluating Collaborative Filtering Recommender Systems, ACM Trans. on Information Systems, Vol.22, No.1, pp. 5-53, 2004.

[ヒリナスキエヴィッチ, 2014] ヒリナスキエヴィッチ イアン, 新谷洋子: Scientific Data データの再利用を促進するオープンアクセス・オープンデータジャーナル, 情報管理, Vol.57, No.9, pp.629-640, 2014.

[Holt & Lampke, 2010] Holt, J.T., and Lampke, E.: Exploreing Stolen Data Markets Online: Products and Market Forces, Criminal Justice Studies, Vol.23, No.1, pp.33-50, 2010.

[Ikeda et al., 1993] Ikeda, M., Kono, Y., Mizoguchi, R.: Nonmonotonic Model Inference -A Formalization of Student Modeling-, 13th International Joint Conference on Artificial Intelligence, pp.467-473, 1993.

[池上ら, 2014] 池上顕真, 早矢仕晃章, 大澤幸生: データ利活用における創造的コミュニケーションと行動プロセス, 創造的社会システムとしてのデータ市場のデザイン, 信学技報, Vol.114, No.343, AI2014-33, pp.45-50, 2014.

[Ikegami & Ohsawa, 2014] Ikegami, K. and Ohsawa, Y.: Modeling of Writing and Thinking Process in Handwriting by Digital Pen Analysis, IEEE-ICDMW 2014, pp.447-454, 2014.

[石井・三輪, 2001] 石井成郎, 三輪和久: 創造的問題解決における協調認知プロセス, 認知科学, Vol.8, No.2, pp.151-168, 2001.

[石井, 2014] 石井夏生利: アメリカのプライバシー保護に関する動向, 情報処理, Vol.55, No.12, pp.1346-1352, 2014.

[和泉, 2003] 和泉潔: 人工市場, 森北出版, 2003.

[Kagel & Roth, 1995] Kagel, J., and Roth, A.: The Handbook of Experimental Economics, Princeton University Press, 1995.

[神嶌, 2006] 神嶌 敏弘: 推薦システム － 情報過多時代をのりきる, 情報の科学と技術, Vol.56, No.10, pp.452-457, 2006.

[神田・石黒, 2000] 神田崇行, 石黒浩: 対話型ヒューマノイドロボットからの日常生活の中の友達関係の推定, 情報処理, Vol.41, No.6, pp.1000-1006, 2000.

[Kandogan et al., 2015] Kandogan, E., Roth, M., Schwarz, P., Hui, J., Terrizzano, I., Christodoulakis, C., and Miller J.R.: LabBook: Metadata-driven Social Collaborative Data Analysis, International Conference on Big Data, pp.431-440, 2015.

[経済産業省, 2009] 経済産業省: 海外におけるオープン・ガバメントの取り組み, <http://www.meti.go.jp/policy/it_policy/e-meti/opengov/opengovreport.pdf>, 2009, 最終アクセス 2016 年 10 月 13 日.

[経済産業省, 2014a] 経済産業省: データ駆動型（ドリブン）イノベーション創出戦略協議会について<http://www.meti.go.jp/press/2014/11/20141105002/20141105002a.pdf>, 2014. [最終アクセス 2016 年 6 月 27 日].

[経済産業省, 2014b] 経済産業省: データ駆動型（ドリブン）イノベーション創出に関する調査事業のためのワークショップ, http://www.meti.go.jp/press/2014/11/20141105002/20141105002.html, [最終アクセス 2016 年 11 月 21 日].

[経済産業省, 2015] 経済産業省: 平成 26 年度経済産業省委託事業, 我が国経済社会の情報化・サービス化に係る基盤整備（データ駆動型イノベーション創出に関する調査

事業）調査報告書. http://www.meti.go.jp/meti_lib/report/2015fy/001102.pdf, 2015 [最終アクセス 2016 年 11 月 21 日].

[経済産業省, 2016] 経済産業省: 平成 26 年度補正経済産業省委託事業, 先端課題に対応したベンチャー事業化支援等事業（データ利活用促進支援事業 : データ駆動型イノベーションを実行するプラットフォーム・プロセス支援）報告書. http://www.meti.go.jp/meti_lib/report/2015fy/001102.pdf, 2016 [最終アクセス 2016 年 11 月 21 日].

[Koga & Aoyama, 2004] Koga T., and Aoyama, K.: Product Behavior and Topological Structure Design System by Step-by-step Decomposition, ASME 2004 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, pp.425-437, 2004.

[古賀・青山, 2010] 古賀毅, 青山和浩: 製品設計における手戻りと矛盾のトレードオフを考慮した設計プロセスの導出手法, 日本機械学会論文集 C 編, Vol.7, No.761, pp.215-224, 2010.

[Kudo & Matsumoto, 2000] Kudo, T., and Matsumoto, Y.: Japanese Dependency Struc-ture Analysis Based on Support Vector Machines, Proceedings of Empirical Methods on Natural Language Processing, pp.18-25, 2000.

[Kushiro et al., 2014] Kushiro, N., Mastuda, S., Torikai, R., and Takahara, K.: A System Design Method Based on Interaction Between Logic and Data Sets, IEEE-ICDMW 2014, pp.462-469, 2014.

[国土交通省, 2016] 国土交通省: ビッグデータの活用手法（IMDJ）の紹介, 国土交通政策研究所報, Vol.61, pp.70-81, 2016.

[Larkin & Simon, 1987] Larkin, J.H. and Simon, H.A.: Why a Diagram is (sometimes) Worth Ten Thousand Words, Cognitive Science, Vol.11, No.1, pp.65-100, 1987.

[Manyika et al., 2013] Manyika, J., Chui, M., Groves, P., Farrell, D., Kuiken, S.V., and Doshi, E.A.: Open data: Unlocking innovation and performance with liquid information, McKinsey Global Institute, 2013.

[Manyika et al., 2015] Manyika, J., Chui, M., Bisson, P., Woetzel, J., Dobbs, R., Bughin, J., and Aharon, D.: The Internet of Things: Mapping the Value beyond the Hype, McKinsey

Global Institute, 2015.

[松原・山本, 1987] 松原仁, 山本和彦: フレーム問題について, 人工知能, Vol.2, No.3, pp.266-272, 1987.

[松尾ら, 2005] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満: Web の情報からの人間関係ネットワークの抽出, 人工知能, Vol.20, No.1, pp.46-56, 2005.

[McDermott & Doyle, 1980] McDermott, D. and Doyle, J.: Non-monotonic logic I, Artificial Intelligence, Vol.13, pp.41-72, 1980.

[Miwa, 2004] Miwa, K.: Collaborative discovery in a simple reasoning task, Cognitive System Research, Vol.5, No.1, pp.41-62, 2004.

[Miyake, 1986] Miyake, N.: Constructive Interaction and the Interactive Process of Understanding, Cognitive Science, Vol.10, No.2, pp.151-177, 1986.

[森ら, 2007] 森裕一, 山本義郎, 宿久洋, 本多啓介: 教育・学習支援のためのデータ指向統計解析環境, 日本統計学会誌, Vol.36, No.2, pp.327-347, 2007.

[森, 2014] 森亮二： 日本の個人情報保護改正の状況，情報処理，Vol.55, No.12，pp.1353-1360, 2014.

[Metcalfe, 1998] Metcalfe, J.S.: Evolutionary Economics and Creative Destruction, Routledge, 1998.

[村上, 1980] 村上陽一郎: 動的世界像としての科学, 新曜社, 1980.

[中崎, 2015] 中崎尚: 個人情報保護法改正とビジネスでの情報利活用への影響, Nextcom, Vol.24, pp.14-25, 2015.

[西村, 2003] 西村行功: シナリオ・シンキング, ダイヤモンド社, 2003.

[西村, 2005] 西村行功: 企業の自己革新におけるシナリオ創発, 人工知能, Vol.20, No.1, pp.9-14, 2005.

[大向, 2013] 大向一輝: オープンデータと Linked Open Data, 情報処理, Vol.54, No.12, pp.1204-1210, 2013.

[Ohsawa et al., 1998] Ohsawa, Y., Benson, N. E., and Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor, In Proc.

Advanced Digital Library Conference (IEEE ADL'98), pp.12-18, 1998.

[大澤, 2003] 大澤幸生: チャンス発見の情報技術, 東京電機大学出版局, 2003.

[Ohsawa & Nishihara, 2012] Ohsawa, Y., and Nishihara, Y.: Innovators' Marketplace: Using Games to Activate and Train Innovators, Springer-Verlag, 2012.

[Ohsawa et al., 2013] 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 22, pp.709-716, 2013.

[大澤, 2014a] 大澤幸生: データジャケット-創造的コミュニケーションのあるデータ市場のために-, 人工知能, Vol.29, No.6, pp.622-627, 2014.

[大澤, 2014b] 大澤幸生: 市場メカニズムを模倣したシステム安全評価に資するデータ共有・活用手法の研究, 平成 25 年度高経年化技術評価高度化事業報告書, 2014.

[大澤ら, 2017] 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（印刷中）

[岡嶋ら, 2015] 岡嶋成司, 山根昇平, 粂照宣: Linked Data 活用を促進するプラットフォーム, 人工知能, Vol.30, No.5, pp.568-573, 2015.

[Rabinovich & Cheon, 2011] Rabinovich, E., and Cheon, S.: Expanding Horizons and Deepening Understanding via the Use of Secondary Data Sources, Journal of Business Logistics, Vol.32, No.4, pp.303-316, 2011.

[Ristoski et al., 2014] Ristoski, P., Mencía, E.L., and Paulheim, H.: A Hybrid Multi-Strategy Recommender System Using Linked Open Data, Semantic Web Evaluation Challenge, Springer, pp.150-156, 2014.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C.S.: A Vector Space Model for Automatic Indexing, Communications of the ACM, Vol.18, No.11, pp.613-620, 1975.

[Salton & Buckley, 1988] Salton, G., and Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval, Information processing and management, Vol.24, No.5, pp.513-523, 1988.

[Sano & Picard, 2013] Sano, A., and Picard, R. W.: Stress Recognition Using Wearable Sensors

and Mobile Phones, Affective Computing and Intelligent Interaction, pp.671-676, 2013.

[佐藤, 2015] 佐藤一郎: パーソナルデータに関わる制度改正動向 : パーソナルデータの利活用と保護の両立に向けて, 電子情報通信学会, Vol.98, No.3, pp.178-187, 2015.

[白松ら, 2016] 白松俊, Teemu Tossavainen, 大囿忠親, 新谷虎松: 社会課題とその解決目標の Linked Open Data 化による目標マッチングサービスの開発, 人工知能, Vol.31, No.1, pp.29-39, 2016.

[Simon, 1955] Simon, H.A.: A Behavioral Model of Rational Choice, The Quarterly Journal of Economics, Vol.69, pp.99-118, 1955.

[砂原ら, 2014] 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.

[諏訪, 1999] 諏訪正樹: ビジュアルな表現と認知プロセス, 可視化情報, Vol.19, No.72, pp.13-18, 1999.

[Sterman, 2000] Sterman, D.J.: Business Dynamics –Systems Thinking and Modeling for a Complex World-, Irwin McGraw-Hill, 2000.

[髙玉, 2010] 髙玉圭樹: コンシェルジュサービスに基づく介護支援システム –パーソナルヘルスデータからライフスタイル設計へ–, 人工知能, Vol.29, No.6, pp.585-590, 2010.

[高崎, 2014] 高崎春夫: 個人情報保護にかかわる法制度をめぐる EU の状況, 情報処理, Vol.55, No.12, pp.1337-1345, 2014.

[武田, 2011] 武田英明: 日本における Linked Data の現状と普及に向けた課題, 情報処理, Vol.52, No.3, pp.326-333, 2011.

[Turney & Pantel, 2010] Turney, P.D., and Pantel, P. 2010. From Frequency to Meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research Vol.37, pp.141-188, 2010.

[Ward & Chapman, 2008] Ward, S., and Chapman, C.: Stakeholders and Uncertainty Management in Projects, Construction Management and Economics, Vol.26, No.6, pp.563–577, 2008.

[Xu et al., 2014]　Xu, L., Jiang, C., Wang, J., Yuan, J., and Ren, Y.: Information Security in Big Data: Privacy and Data Mining, IEEE Access, Vol.2, pp.1149-1176, IEEE, 2014.

[山崎・三輪, 2001]　山崎治, 三輪和久: 外化による問題解決過程の変容, 認知科学, Vol.8, No.1, pp.103-116, 2001.

[Yu, 2011]　Yu, L.: A Developer's Guide to the Semantic Web, Springer, 2011.

[是津, 2007]　是津耕司: メタデータ検索技術の現状と今後の展望, 映像メディア学会誌, Vol.61, No.2, pp.146-151, 2007.

公表済み研究成果

1. 査読付きの学術論文誌（ジャーナル）：6 報

- Teruaki Hayashi, Yukio Ohsawa, “Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game plus Action Planning,” International Journal of Knowledge and Systems Science, Vol.4, No.3, pp.14-38, 2013.
- Teruaki Hayashi, Yukio Ohsawa, “Shikakeological Approach of Innovators Marketplace as Role-based Game and Evaluation Method for Solutions,” AI & Society, Vol.30, No.4, pp.451-461, Springer London, 2014.
- Teruaki Hayashi, Yukio Ohsawa, “Relationship between Externalized Knowledge and Evaluation in the Process of Creating Strategic Scenarios,” Open Journal of Information Systems (OJIS), Vol.2, No.1, pp.29-40, 2015.
- Yukio Ohsawa, Teruaki Hayashi, “Tangled String for Sequence Visualization as Fruit of Ideas in Innovators Marketplace on Data Jackets,” Intelligent Decision Technologies, pp.1-13, 2016.
- Teruaki Hayashi, Yukio Ohsawa, “Comparison of Conflict Resolution Behavior and Scenario Generating Process in Group and Individual by Handwriting Process Analysis,” Intelligent Decision Technologies, pp.1-9, 2016.
- 早矢仕晃章, 大澤幸生, “データ利活用知識構造化と再利用による検索システム: Data Jacket Store,” 人工知能学会論文誌, Vol.31, No.5, 2016.

2. 書籍・特集・解説論文：6 報

- Yukio Ohsawa, Hiroyuki Kido, Teruaki Hayashi, Chang Liu and Kazuhiro Komoda, “Innovators Marketplace on Data Jackets, for Valuating, Sharing, and Synthesizing Data,” Knowledge-based Information Systems in Practice, Smart Innovation, Systems and Technologies, Jeffrey. W., Tweedale, Lakhmi. C. Jain, Junzo Watada and Robert Howlett (eds), Springer-Verlag, Vol.30, pp.83-97, 2015.
- 早矢仕晃章, 大澤幸生, “データジャケットを用いた市場型ワークショップ（IMDJ）とその活動状況,” ヒューマンインターフェース学会誌, データジャケットを用いた市場型ワークショップの展開～産学官によるデータ利活用価値の発見～特集, Vol.17, No.2, pp.107-114, 2015.
- 秋元正博, 高階勇人, 野深裕也, 大澤幸生, 早矢仕晃章, “データ市場の創生,” 統計, 特集 ビッグデータ, Vol.66, No.9, pp.26-31, 2015.

- 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（印刷中）うち以下の 2 つは分担執筆
  - 早矢仕晃章, “データ市場,” 大澤幸生著の書籍分担：第 4 章「データジャケット」, 近代科学社, 2017.（印刷中）
  - 早矢仕晃章, “データ市場,” 大澤幸生著の書籍分担：第 5 章「アクション・プランニング」, 近代科学社, 2017. （印刷中）
- Yukio Ohsawa, Teruaki Hayashi, Hiroyuki Kido, “Restructuring Incomplete Models in Innovators Marketplace on Data Jackets,” Chapter 48, Springer Handbook of Model-Based Sciences, Lorenzo, M., and Tommaso, B. (eds), Springer International Publishing, 2017. （印刷中）

3. 査読付きの国際会議：14 報

- Teruaki Hayashi, Yukio Ohsawa, “Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game plus Action Planning,” 20th European Conference on Artificial Intelligence, 1st European Workshop on Chance Discovery and Data Synthesis in ECAI 2012, Aug. 2012.
- Yukio Ohsawa, Hiroyuki Kido, Teruaki Hayashi, Chang Liu, “Data Jackets for Synthesizing Values in the Market of Data,” 17th International Conference in Knowledge Based and Intelligent Information and Engineering System (KES2013), pp.709-716, Procedia Computer Science, Vol.22, 2013.
- Yukio Ohsawa, Hiroyuki Kido, Teruaki Hayashi, Masahiro Akimoto, Masanori Fujimoto, Masaki Tamada, “Strategies for Creative Argumentation: Learned from Logs of Innovators Market Game,” Proc. of the 2013 International Conference on Brain & Health Informatics (BHI'13), LNCS, Oct. 2013.
- Yukio Ohsawa, Teruaki Hayashi, “Workshop with Tsugo Roulette for Externalizing Intentions and Constraints of Participants in Social Activities,” The Third International Workshop on Advanced Computational Intelligence and Intelligent Informatics, Oct. 2013.
- Yukio Ohsawa, Chang Liu, Teruaki Hayashi, Hiryuki Kido, "Data Jackets for Externalizing Use Value of Hidden Datasets," 18th International Conference on Knowledge Based and Intelligent Information and Engineering System (KES2014), pp.946-953, Procedia Computer Science 35, 2014. DOI: 10.1016/j.procs.2014.08.172
- Teruaki Hayashi, Yukio Ohsawa, “Estimation of Novelty Assessment of Strategic Scenarios Using Relativeness,” IEEE-ICDM Workshops 2014, Designing the Market of Data - for

Practical Data Sharing via Educational and Innovative Communications (MoDAT), pp.411-446, Dec. 2014.

- Teruaki Hayashi, Yukio Ohsawa, "Knowledge Structuring and Reuse System Design Using RDF for Creating a Market of Data," 2nd International Conference on Signal Processing and Integrated Networks (SPIN2015), pp.607-612, Feb. 2015.
- Teruaki Hayashi, Yukio Ohsawa, "Knowledge Structuring and Reuse System Using RDF for Supporting Scenario Generation," 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES2015), Procedia Computer Science, Vol.60, pp.1281-1288, 2015.
- Yukio Ohsawa, Teruaki Hayashi, "Visualizing History for Qualitative Explanation of Valuable Events using Tangled String," 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES2015), Procedia Computer Science, Vol.60, pp.1178-1185, 2015.
- Teruaki Hayashi, "Estimating Contextual Relationships of Stakeholders in Scenarios Using DBpedia," IEEE-ICDM Workshops 2015, Designing Safe and Secure Life on the Market of Data -Exchanging and Integrating Data via Insightful Communications- (MoDAT), pp.698-705, Nov. 2015. DOI: 10.1109/ICDMW.2015.16
- Teruaki Hayashi, Yukio Ohsawa, "Meta-data Generation of Analysis Tools and Connection with Structured Meta-data of Datasets," 3rd International Conference on Signal Processing and Integrated Networks (SPIN2016), pp.226-231, Delhi, India, Feb. 2016. DOI: 10.1109/SPIN.2016.7566693
- Teruaki Hayashi, Yukio Ohsawa, "Preliminary Case Study about Analysis Scenarios and Actual Data Analysis in the Market of Data," 2nd European Workshop on Chance Discovery and Data Synthesis (EWCDDS16) in ECAI 2016, Sep, the Hague, Netherlands. 2016.
- Teruaki Hayashi, Yukio Ohsawa, "Comparison between Utility Expectation of Public and Private Data in the Market of Data," 20th International Conference on Knowledge Based and Intelligent Information and Engineering System (KES2016), Vol.96, pp.1267–1274, Sep, York, UK. 2016. DOI: 10.1016/j.procs.2016.08.171
- Teruaki Hayashi, "Inferring Variable Labels Considering Co-occurrence of Variable Labels in Data Jackets," IEEE-ICDM Workshops 2016, Data Market for Co-evolution of Sciences and Business (MoDAT), pp.783-787, Dec. 2016. DOI: 10.1109/ICDMW.2016.11

4. 査読のない国際会議（Invited Paper）：1 報

・ Teruaki Hayashi, Yukio Ohsawa, “The Efficient Method for Creating Ideas: Innovators Marketplace as Role-Based Game,” Shikakeology: Designing Triggers for Behavior Change: Papers from the 2013 AAAI Spring Symposium, pp. 27-32, Mar. 2013.

5. 査読付き国内会議：1 報

・ 早矢仕晃章, 大澤幸生, “筆記行動分析に見る集団と個人のシナリオ生成プロセスと矛盾解消行動の比較,” 第 14 回情報科学技術フォーラム（FIT2015）, pp.51-56, Sep. 2015.

6. 査読のない国内会議（研究会など）：16 報

・ 早矢仕晃章, 大澤幸生, “制約とコミュニケーションによるアイデア精緻化メソッド：アクション・プランニング,” 第 12 回情報科学技術フォーラム(FIT 2013), pp. 405-408, Nov. 2013.
・ 早矢仕晃章, 大澤幸生, “データの価値と利用方法発見のための創造的コミュニケーションとメタデータ記述方法の提案,” 2014 年度人工知能学会全国大会（第 28 回）, May. 2014.
・ 大澤 幸生, 平岡 美那子, 野神 航一, 劉 暢, 早矢仕 晃章, “Tangled String: データジャケットを用いたイノベーションゲームからのデータ可視化手法創成,” 電子情報通信学会「人工知能と知識処理」(AI)研究会, SIG-SWO-A1401-04, Aug. 2014.
・ 早矢仕晃章, 大澤幸生, “協同問題解決のプランニングにおける発言とシナリオ評価への影響,” 電子情報通信学会ヒューマンコミュニケーション基礎研究会（HCS）, 信学技報, Vol.114, No.273, pp.37-42, Oct. 2014.
・ 池上顕真, 早矢仕晃章, 大澤幸生, “データ利活用における創造的コミュニケーションと行動プロセス,” 創造的社会システムとしてのデータ市場のデザイン, 信学技報, Vol.114, No.343, AI2014-33, pp.45-50, Nov. 2014.
・ 早矢仕晃章, 大澤幸生, “データ市場創造のための技術及び知識再利用法の提案,” 第 6 回システム創成学学術講演会, Dec. 2014.
・ 早矢仕晃章, 大澤幸生, “データ利活用知識の構造化と再利用によるデータ情報推薦システム Data Jacket Store の提案,” 電子情報通信学会 ライフインテリジェンスとオフィス情報システム研究会（LOIS）, 信学技報, Vol.114, No.500, pp.61-66, Mar. 2015.
・ 早矢仕晃章, 大澤幸生, “データ市場創造のための知識再利用と実行動を促すシナリオ生成手法の提案,” 情報処理学会第 77 回全国大会, pp.39-40, Mar. 2015.
・ 早矢仕晃章, 大澤幸生, “データ利活用知識検索システム DJ Store を用いた利用価値の

あるデータの特徴抽出," 2015 年度人工知能学会全国大会（第 29 回）, May. 2015.

- 張確軒, 早矢仕晃章, 大澤幸生, "日本語版 DBpedia によるアノテーションを介した要求と解決策の意味的関係の可視化," 2015 年度人工知能学会全国大会（第 29 回）, May. 2015.
- 早矢仕晃章, 大澤幸生, "DBpedia データセットを用いたステークホルダー表出とシナリオにおける関係推定," 第 36 回 SWO 研究会-DBpedia シンポジウム, SIG-SWO-036-06, Jul. 2015.
- 早矢仕晃章, Randy Goebel, 大澤幸生, "変数名と分析ツールの顕在化を促進する擬似コード形式記述を用いた分析シナリオ生成手法の提案," 第 7 回システム創成学学術講演会, Jan. 2016.
- 早矢仕晃章, 大澤幸生, "データジャケットにおける変数ラベルの共起性を考慮した変数ラベル推定手法の検討," 2016 年度人工知能学会全国大会（第 30 回）, 福岡, Jul. 2016.
- 山内雄太, 村岡恒輝, 野中尚暉, 早矢仕晃章, 今泉允聡, 金善右, 松並研作, Lee H.W., 鎗目雅, "書誌情報を用いた多分野にわたる研究インパクトの評価法の検討," 電子情報通信学会ソサエティ大会 2016, 北海道, Sep. 2016.
- 早矢仕晃章, 大澤幸生, "実装的データ市場における分析シナリオを用いたデータ利活用プロセスの実験的考察," 電子情報通信学会 人工知能と知識処理研究会（AI）: データ市場特集 III, 岡山, Feb. 2017.
- 村岡恒輝, 山内雄太, 野中尚暉, 早矢仕晃章, 今泉允聡, 金善右, 松並研作, Lee H.W., 鎗目雅, "多分野にわたる研究インパクトの評価法の検討," 電子情報通信学会 人工知能と知識処理研究会（AI）: データ市場特集 III, 岡山, Feb. 2017.

講演

- 「Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game plus Action Planning」パリ第 6 大学（UPMC-LIP6）招待講演（2012 年 8 月・フランス）
- 「イノベーションゲームワークショップ」招待講演（2013 年 5 月・神奈川:横浜市立大学 学術院国際総合科学群）
- 「チャンス発見とバングラデシュから学ぶシナリオ・デザイン」チャンス発見学 15 周年記念シンポジウム記念講演・基調講演（2014 年 8 月・東京）
- 「データ・ジャケット：データ市場創造のために技術～データの価値発見からビジネス創出へ～」オープンデータ／ビッグデータ利用推進フォーラム 第 1 回利活用セミナー・基調講演（2014 年 9 月・大阪）

- 「データジャケット：データ市場創造のための技術」日経 BP コンソーシアム招待講演（2014 年 11 月・東京）
- 「データ駆動型イノベーション創出に関する調査事業のためのワークショップ：アクション・プランニングについて」経済産業省の委託事業「我が国経済社会の情報化・サービス化に係る基盤整備（データ駆動型イノベーション創出に関する調査事業）講演（2014 年 12 月・東京）
- 「データ市場創造技術」情報銀行コンソーシアム招待講演（2016 年 10 月・東京）

研究助成金

- ¥247,303, IEEE International Conference on Data Mining Workshop（中国）への参加に関する支援, 東京大学リーディング大学院プログラム GSDM（Global Leader Program for Social Design and Management）, 短期派遣プログラム, 2014.
- ¥1,000,000, 研究課題「データ市場活性化を促すデータ可視化手法の開発」, 東京大学リーディング大学院プログラム GSDM（Global Leader Program for Social Design and Management）, University of Alberta（Edmonton, Canada）における国際プロジェクト演習「共同研究」, 2015.
- ¥700,000, 研究課題「データ利活用知識構造化と再利用による意思決定支援システムの研究」, 平成 28 年度科学研究費助成事業（特別研究員奨励費）, 2016.
- ¥297,472, IEEE International Conference on Data Mining Workshop（スペイン）への参加に関する支援, 東京大学リーディング大学院プログラム GSDM（Global Leader Program for Social Design and Management）, 短期派遣プログラム, 2016.