Key Perspectives

Consultants in scholarly information

データ科学者とキュレーターの スキル、役割、キャリア構造: 現在の業務と今後のニーズの評価

JISC への報告書

2008 年 7 月

作成:
アルマ・スワン、シェリダン・ブラウン
キー・パースペクティブ社
48 Old Coach Road
Playing Place
Truro, TR3 6ET
UK
+44 1392 879702
aswan@keyperspectives.co.uk
www.keyperspectives.co.uk

目次

1. 要旨

本研究は、英国におけるデータ管理に関する Liz Lyon による報告書（Lyon, 2007）で提言された 2 つの勧告に具体的に対処するために JISC により委託されものである。本研究の主な目的は、データ科学者の役割とキャリア形成、関連する専門的データキュレーションスキルの研究コミュニティへの供給に関して調査し、勧告を行うことであった。

現在使用されている用語は不正確であり、データ関連の既存の様々な役割に関して誤解を与える可能性がある。我々は 3.1 節において、権威ある組織が与えた定義とこの分野で働く人々の実務経験を調整しようと試みた。我々は次の 4 つの役割、データ作成者、データ科学者、データ管理者、データ図書館員を区別して、以下のように簡単に定義する。

- データ作成者: 該当分野の専門知識を持ちデータを生産する研究者。これらの人々は、データを処理、操作、使用する高水準の専門知識を持っていると思われる。
- データ科学者: 研究が行われている場所で働く（データセンターの職員の場合は、データ作成者と密接に協力して働く）人々。研究者がデジタルデータを使って研究が行えるようにするために、創造的な調査や分析に従事し、データベース技術の開発を行う。
- データ管理者: コンピュータ科学者、情報技術者、情報科学者のいずれかで、コンピュータ機器、ストレージ、持続的アクセス、データ保存に責任を有する人々。
- データ図書館員: 図書館コミュニティを出自とし、データのキュレーション、保存、保管を行う教育を受け、これらを専門に行う者。

現実的には、今のところこれらの用語はデータコミュニティにおいて正確には使用されておらず、各役割の境界もあいまいであると思われる。正確な用語が一般に浸透するには時間がかかるだろう。

データ科学は、現在国際的に注目を集めている話題であり、米国、カナダ、オーストラリア、英国、欧州において、進展が見られている。データ科学に関してはこれらすべての国において、各課題に個別的に対処するのではなく、国家的規範に基づいて組織化され、発展されるべきであると考えられていることは注目に値する。

一般に、研究者はデータを基本とする研究がもたらす問題を以前よりずっと意識するようになっている。既にデータを処理・管理する多大なスキルを持つ研究者（いわゆる「ネイティブ・データ科学者」）もいるが、このことについてあまり経験のない研究者もさらに学習することに興味を示している。周囲にデータ科学者がいない研究者は、所属機関の IT サービス部門や図書館に支援や

アドバイスを求めている。英国では、現在、一般的なデータスキルの向上を図ることを目的とするデータ管理を教える修士課程を提供する大学が出始めている。データセンターがデータ科学者をかなり長い時間をかけて教育したにもかかわらず結局は他の仕事を求めて辞めていくのを受け入れ、それによりデータスキルが研究コミュニティに普及することを助けたように、特にデータ関連の大学院教育を受けた研究者の数が増えることで同じことが行われるだろう。

通常、データ科学者は計画的ではなく偶然の結果として現在の職業についている。しかし、データ科学のポストが増えるにつれ、状況は変わっている。データ科学者がその職務に就くには、その分野の専門家がそのキャリアの中で専門的なデータスキルを習得するか、コンピュータ科学者を出自とする者が時間をかけて専門分野の知識を習得するかのいずれかである。現在データ科学者のポストにいるほとんどの者は、自らのスキルは仕事を通じて習得したと述べている。適当な研修機会がないことやイベントに参加するためのコスト（時間とお金）のためである。最近までは、資格に関する明確な決まりが存在していなかったが、現在では、情報学に関する大学院レベルの教育が必要とされることが多くなっている。実際、データ科学者は幅広いスキルを必要とする。専門分野の知識とコンピュータスキルは必須条件であるが、「対人スキル」も重要である。なぜなら、その主要な職務の 1 つは、研究者のニーズや実務を（我々がデータ管理者と定義した）コンピュータ専門家に、また場合によってはその逆に、翻訳することであるからである。

データ科学者には確定したキャリア構造が存在しない。これは、英国の研究コミュニティにデータスキルが適切に供給されるようになるためには解決されなければならない主要な問題である。データ科学者は大学やデータセンターに終身職として採用されている場合もあれば、短期的な研究契約で雇用されている場合もある。大学において終身職を持つ者は、サービス職または教員関連職から完全な教員職まで様々な種類の職階にある。現在はシステム全般にわたる一貫性は存在しない。職が安定していないことは、データ科学者になることを勧めたり教育したりする際の問題であり、現在は、スキルを持つ者への需要は供給を大きく上回っている。データ科学者（やその予備軍）の不満の原因となっているもう 1 つの問題は、彼らの役割が専門化されていないことや正式で組織化されたキャリア構造がないことにより、過小評価されていると感じることである。

データ科学の職務に携わる人は、適切なスキルアップを続けるという持続的で大きな課題に直面している。データに関する事柄は急速に変化しているので、データ科学者は一般的な進展や自らの専門分野固有の進展に常についていく必要がある。これを支援する国際的ワークショップが存在する分野もあるが、たとえそうであっても、常に十分であるというわけではない。データ科学者は「今現在最も重要な」特定のトピックに関する短期コースを定期的に受講するという形の継続的専門性開発という考えを好んでおり、そのようなシステムが一般に認められた職務の一部となることを希望している。

学部教育カリキュラムにデータスキルに関する科目を取り入れることに価値があるか否かという質問に関しては意見が 2 つに分かれた。多くの人はこれには価値があると考えており、データ科学者も教育機会が早ければ早いほど、将来の研究者は基本的データスキルをより良く身に付けることができると考えている。一方、学部教育を行っている多くの者は既に講義は手一杯であり、特別なデータスキル科目を追加できないと述べている。さらに、データ処理スキルが特に必要な分野においては、(簡単なリレーショナルデータベースの構築法や使用法を教えているなど)既に学部カリキュラムに取り入れられていると指摘した。今後事態が進展すれば、当然、各専門分野に相応しい方法によるさらなるデータスキル教育が学部教育に取り入れられるようになるだろう。

データ量の多い研究に対する図書館の役割は重要であり、今や研究支援に関する図書館の立ち位置を戦略的に変更すべき時である。我々は図書館には主な潜在的役割が 3 つあると考えている。第 1 に研究者の間にデータに対する認識を増加させることであり、第 2 に機関リポジトリを通じて機関で生産されたデータを保管・保存するサービスを提供することであり、第3に図書館のデータサービスに対する一連の新しい専門的業務を開発することである。米国では、この点に関して既に進展が見られる。図書館コミュニティはデータの氾濫という需要に合わせて、ライブラリスクールにおける教育を通じて、データを保管・保存するスキルを正式に提供する体制を構築した。英国においてもこの分野における進展の萌芽が見られる。しかし、未だ専門家としてのデータ図書館員の数は十分ではない。英国における現時点の数は 5 名に過ぎず、この状況は迅速に変える必要があると考えられる。現在そのように数が少ない理由の 1 つは、データ科学者の状況と同じであり、一般に認められているキャリアパスがないことである。資質の高い人、すなわち、特定の分野の経験があり、当然その分野におけるデータの構造や使用法を理解している人を惹きつけることも現時点では困難である。また、英国よりもはるかに進んでいる米国では、データ図書館員研修生のための適当な実務研修先がないことが専門家としての教育を妨げる要因として特定されている。これはやがて英国でも問題になると思われる。

本研究による主な勧告は次の通りである。

1. 研究ドメインにおけるデータスキル開発に関する勧告(RD)

勧告 RD1: 英国の主要な研究助成団体は、大学および研究機関と協力して、データ科学者の役割を正しく定義および形式化し、データ科学者の仕事を認識させ、報いることができる方法を策定するべきである。

勧告 RD2: 上と同じ団体は協力して、データ科学を支援し、その研究を促進し、その職務の専門化を促進する条件を作成するべきである。

勧告 RD3: JISC および研究を委託するその他の組織は、次の課題を扱う研究を推進するべきである。

- データ科学者により担われる役割およびデータ科学者が研究に果たす貢献の価値の記述
- データ科学キャリアの例
- データ科学におけるグッドプラクティスを表す一連の実務の開発

勧告 RD4: 関連団体(HEFCE と研究会議)は、データ管理の基礎を扱いそれにより基本的なデータ科学スキルを研究プロセスに組み込む、大学院レベルの短期研修コースを研究者に提供するスキルを持つ指導者のネットワークの構築とそれへの資金提供を検討するべきである。既にいくつかの研究会議は、助成金交付申請にデータ計画の記入を要件とすることで、このための基礎を構築している。

勧告 RD5: 研究会議およびその他の研究助成団体は、助成金交付申請および交付決定プロセスの一部として、プロジェクトチームのうち少なくとも 1 人をプロジェクトのデータ科学者として指名することを要件とするべきか否かを検討するべきである。指名された者にはデータ科学とデータ管理の基礎を提供する短期コースに参加することを必須とするべきである。研究会議は、正当なコースの認定と出席証明が必要とされる範囲を検討するべきである。

2. 研究図書館におけるデータスキル開発に関する勧告(RL)

勧告 RL1: 英国の研究図書館コミュニティは、大学および研究機関と協力して、データ図書館員の役割を正しく定義および形式化し、データ処理スキルを持つ図書館員の適切な供給を保障するカリキュラムを開発するべきである。

勧告 RL2: JISC は、国際データキュレーション教育活動(IDEA)作業グループの発展を支援することを検討するべきである。このグループは、特に、図書館・情報学を出自とする将来のデータ図書館員のための適当なカリキュラムの作成において重要な諮問的役割を果たす位置にいる。

3. 一般的なデータスキル開発に関する勧告(RG)

勧告 RG1: 既に、データ分野で活動している多くの関係者が存在するので、データスキル研修において相乗効果を有効に活用できる可能性がある。この可能性、特に、UK データアーカイブ、データ科学分野を先導する大学や研究グループ、ライブラリスクール、デジタルキュレーションセンター、IDEA(国際データキュレーション教育連合)の活動を視野に入れた研究を勧告する。この研究は、米国やカナダ、オーストラリアの活動など国際的な調査も行った方が良いだろう。

2. 序論および方法論

「サイバーインフラストラクチャにより切り開かれた第5の次元は、デジタル時代を革新する新しい力である… この第 5 の次元を受け入れない個人、グループ、組織、国はデジタル時代に取り残されるであろう」(Christopher Greer[1], 2007)

本研究は、デジタル研究データに関する英国の概況を示した報告書「データ処理: 役割、権利、責任、関係」(Lyon, 2007)で提言された 2 つの勧告を実行するために、JISC により委託されものである。数多くの勧告の中から本報告書で取り上げるのは次の 2 つの勧告である。

勧告 34。データ科学者の役割とキャリア開発、および、専門家としてのデータキュレーション技術の研究コミュニティへの提供を調査する研究が必要である。

勧告 35。JISC は、データを処理、キュレーション、保存するスキルを大学や大学院のカリキュラムに取り入れる価値と可能性を評価する研究に助成するべきである。

これらの勧告は、あらゆる専門分野の研究からデジタルデータがあふれるように吐き出されている「データの氾濫」状況においてなされたものである。何故データの問題に注意を払わないといけないのかという疑問に対する理由としては、いわゆる「ビッグサイエンス」や「e サイエンス」の存在が常に挙げられる。もちろん、ビッグサイエンスが大量のデータを生み出すのは事実である。全研究データのおよそ 80%は、高エネルギー物理学、気象学、天文学の 3 分野で生産されている。とはいえ、「スモールサイエンス」もデータ氾濫の一翼を担っており、その貢献度も増加している。これらすべてのデータを管理し、ケアする必要がある。これらのデータを(通常その作成者には想像できない形で)再利用可能にする技術は、研究の進歩にとって非常に重要である。データ科学者は、このようなデータの処理、操作、キュレーションスキルを研究コミュニティに提供し、データ図書館員は、データ成果物を安全に管理することを保障する保管・保存スキルを提供する。

データ科学とデータケアの重要性は、研究データをロングテールの観点から考えるとさらに明らかになる。いわゆるビッグサイエンスプロジェクトで生産されるデータは*比較的*均一であり、技術的観点から言えば取り扱いが容易である場合が多い。一方、スモールサイエンスのデータ成果物は、きわめて不均一であり、データの作成や処理に独特な方法を必要とする場合が多い。手短に言えば、ロングテールに属するデータは取り扱いや再利用、保存が難しいが、大きな潜在的価値を持っている。また、スモールサイエンスにおけるデータの作成は確実に高くつく。全米科学財団が 2007 年に生物学分野の研究に給付した研究費を分析した最近の調査は、総給付研究費の 44%がスモールサイエンスプロジェクト(給

[1] Christopher Greer, 全米科学財団サイバーインフラストラクチャ室

付研究費が 35 万ドル以下のプロジェクト）であることを示した[2]。スモールサイエンスプロジェクトにより生産されるデータを再利用する潜在的価値を解明する方法を提供することは、データ科学者とデータ図書館員が立ち向かうべき重要な課題である。

方法論

本プロジェクトでは、人々の考え方を詳細に調査できるようにするために、質的方法を重視した多面的アプローチを採用した。まず、半定型的で詳細なインタビューを 57 名に対して行い、一方で、データ科学者（研究グループに参加している者とデータセンターや研究会議に勤務している者を含む）、図書館員、図書館技術者、教育者の意見を聞くために 4 つのフォーカスグループを作成した。フォーカスグループとインタビューは、イングランド、北アイルランド、スコットランドで実施した。そして、システム生物学、天文学、化学、考古学、地質学、生態学、地域経済学、土地利用学、社会科学の諸分野を含む幅広い様々な専門分野の研究者の意見を求めた。さらにこのプロセスは、データ科学者に対するオンライン調査と徹底した机上調査による裏付けが行われた。また、米国と英国の専門家が参加してデジタルキュレーションカリキュラムの開発を議論するワシントン DC で開催された 2 日間のワークショップと、データ関連の問題を研究している JISC 助成の様々なプロジェクトの開かれた交流を促進する JISC 主催の会議に参加した。

調査結果をおよそ 30 ページの明快な文書にまとめることと勧告数を最大 10 に抑えることは本プロジェクトの課題の 1 つであった。

本研究のために寛大にも時間を割き、意見をいただいたすべての方に感謝いたします。これらの人々は皆忙しいにもかかわらず、本研究のために喜んで参加してくださいました。

アルマ・スワン、シェリダン・ブラウン
キー・パースペクティブ社
英国トルロー
2008 年 9 月 1 日

[2] BP Heidorn, Graduate School of Library and Information Science University of Illinois at Urbana-Champaign & the NSF, June 2008 （個人的連絡による）

3. データ科学問題の概要

3.1 諸定義

本プロジェクト開始後すぐに、何をする誰を何と呼ぶかという一般的な用語の使用法が未だ整備されていない問題が明らかになった。本プロジェクトのスポンサーは研究の対象となる役割に対し「データ科学者」という用語を使用し、その役割の内容を*データの処理*、*キュレーション*、*保存*に関する仕事であるとした。しかし、自らをデータ科学者だと考えている者は、これらすべての仕事を行うかもしれないが、中でも最初の仕事、すなわち、データ処理に最大の重点を置いており、*必ずしも*自らをデータキュレータやデータ保存者とは考えていないことが本研究で明らかになった。多くの場合、これらの役割は高度な専門性を持つ者により実行される別個のものである。

また、2 つのスペクトルが存在する。1 つは、デジタルデータを使った、あるいはデジタルデータに関して行われる仕事のスペクトルであり、今 1 つは、人々が受け入れている職務の名称のスペクトルである。前者は 4.1 節で取り上げることにし、ここでは、本研究が対象とする役割に名前をつける何らかの実行可能な結論を導く。

全米科学財団の*科学委員会*は 2005 年に公開された報告書(NSF, 2005)で 1 つの定義を提出した。報告書では次のようなデータ科学者の創出を求めている。

> *「デジタルデータコレクションの管理を成功させるために不可欠な情報およびコンピュータ科学者、データベースおよびソフトウェア技術者、プログラマ、各学問分野の専門家、キュレーター、専門注釈家、図書館員、文書館員など」*

これは、コンピュータ科学者から図書館コミュニティのメンバーに至る、純粋に研究上の役割を持つ人々を包括的に示したものである。報告者は続けて、これらの人々は次のようなことを行うと述べている。

> *「創造的な調査と分析を行い；協議、協力、調整を通じてデジタルデータコレクションを使って研究や教育を行う他の人々の能力を高め；データ視覚化法や情報発見法などのデータベース技術や情報科学の革新的概念の開発を行い、これらをコレクションに関連する科学や教育分野に適用する最前線に立ち；最善の方法と技術を実施し；新たに研究を始める、あるいは別の分野から移行してきた研究者、学生、およびデータ科学の追及に関心を持つその他の人々の指導者となり；データコレクションやデジタル情報科学の利益を、考えうる最大範囲の研究者、教育者、学生、一般大衆が利用できるようにするための教育やアウトリーチプログラムを計画・実施する。」*

報告書はさらに、データ科学者の役割と次の 3 つの役割を区別している。

- *データ著者: 科学者、教育者、学生、その他デジタルデータを生産する研究に参加する人々。これにはデータから生み出される研究に常に関心を持っている専門科学者、教育者、学生を含む。*
- *データ管理者: データベースの運用と保守に責任を持つ組織および**データ科学者**(混乱を生じさせる可能性があることを示すために強調した)とデータの保管・保存における信頼できる優秀なパートナー。*
- *データ利用者: 幅広い研究・教育コミュニティ。それらを代表する専門的・科学的コミュニティを含む。*

英国の研究コミュニティにおけるデータ管理の実務、特に、プロジェクトスポンサーにより要求のあった高等教育機関における実務を研究した結果、我々はこのような区別を採用しないことにした。我々の意見では、各役割は次のように明確に区別される。

- データ作成者またはデータ著者: 該当分野の専門知識を持ちデータを生産する研究者。これらの人々は、経験を通して得た、または、必要に迫られて、または個人的関心の結果として得た、データを処理、操作、使用する高水準の専門知識を持っていると思われる。
- データ科学者: 研究が行われている場所で働く(データセンターの職員の場合は、データ作成者と密接に協力して働く)上に示した NSF の定義に記述されているすべての、または多くの機能を果たす人々。多くの場合、自らがデータ作成者である場合も含む。データ科学者はその出自と教育において、担当分野の専門家、コンピュータ科学者、または情報技術者の場合があるが、そのキャリア形成において自らの専門分野ではない分野のスキルを吸収する必要があったと思われる。そのため、例えば、あるシステム生物学分野のデータ科学者は、大量のコンピュータスキルを習得した生物学者であるかもしれないし、大量の生物学知識を習得したソフトウェア工学者かもしれない。自らの重要な役割の中には、データ作成者のニーズをデータ管理者に伝える「通訳」になること(以下を参照)や、データ管理者と共に作業して、データを利用できる形で保管しアクセスできるようにすることだと述べたデータ科学者もいた。
- データ管理者: コンピュータ科学者、情報技術者、情報科学者のいずれかで、コンピュータ機器、ストレージ、持続的アクセス、データ保存に責任を有する人々。データ科学者と密接に連携して、研究グループが各自の研究を効率的に行えるように、常に適切な機器を利用可能な状態にしておく。自らの役割は、データをある場所から別の場所へ送ることであり、常にデータが正確に流れるようにし、また、貴重なデータが失われないようにするデータの「配管工」だと述べたデータ管理者もいた。
- データ図書館員: 図書館コミュニティを出自とし、データのキュレーション、保存、保管を行う教育を受け、これらを専門に行う者。元々、データ図書館員という用語は社会科

学のデータを扱う図書館員に制限されるものだと考えられていたが、現在では、この肩書きはデータスキルを持つ全分野の図書館員を示している。研究成果を収集・管理するデジタルリポジトリの構築を開始する機関には特に重要な役割である。データセットは研究成果の一部であり、機関リポジトリはこのようなデータの自然な所蔵場所であり、リポジトリは通常図書館で運営されている。「ビッグサイエンス」は各自の（国際的な）データセンターを持っており、英国ではいくつかの研究会議が全国データ格納施設を提供しているが、「スモールサイエンス」についてはそのような施設が各機関で提供される必要があると思われる。たとえ全国データサービスという形の第 3 のプレイヤーが実現したとしても、何らかの理由で全国データセンターに保管される資格が得られないデータのためにローカル施設が必要になるだろう。データ図書館員はそのような大量データの管理人となる。

現実的には、これらの用語がここで定義した通りに使用されていないことは認識している。我々がデータ科学者だと考える人々はデータ管理者またはデータ専門家と呼ばれる職業についている。用語の標準化には時間がかかるだろうが、本報告書ではこの定義に従うことにする。なお、現在のところ、これらの役割の境界は極めてあいまいであることに注意することも重要である。今のところ、これら役割を発展させる主な推進力は実用主義、ニーズ、個人の好みである。データ科学者の役割がより一般的になり、機関がデータ成果物に対する新たな責任を認識するようになれば、研究コミュニティや図書館がデータ管理に対して共通のアプローチをとらざるを得ないようになり、何らかの変化が生ずることが期待される。

3.2 各国のアプローチ

データを管理する人がどんな名前で呼ばれているにせよ、これらの人々に対する政府や幅広い研究コミュニティの関心は益々高くなっている。ここではいくつかの国の概要を簡単に報告する。

カナダでは、（UKDA のような）英国モデルに基づく全国的データアーカイブシステムは存在せず、政府の一部も新機関の創設に乗り気でないが、データ戦略に関してはその責任を果たすべき既存機関を探すという問題になってきている。研究会議が協議を行い、カナダ国立図書館・文書館（国立図書館と国立文書館が合併）が 18 ヶ月の研究を行い、報告書を発表した（Canadian Digital Information Strategy, 2007）。CARL（カナダ研究図書館協会）も、CISTI（カナダ科学技術情報研究所）によりコーディネートされた研究データカナダ（Research Data Canada）と呼ばれるタスフォースと共に、データ戦略と機関リポジトリを結びつけるという観点から、この問題に注目している。

オーストラリア政府は、オーストラリア全国データサービス（ANDS: Australian National Data Service – http://ands.org.au/）を構築し、2008 年最終四半期には本運用する予定である。このサービスは 4 つの主要な活動プログラムを持つ。開発フレームワーク（Developing Frameworks）

プログラムは国家的政策と機関ポリシーを扱い、ユーティリティ提供（Providing Utilities）プログラムは発見サービスと機械間サービスを扱い、コモンズ育成（Seeding the Commons）プログラムはオーストラリアにおける研究のためのデータコモンズを構築し、可能性開発（Developing Capabilities）プログラムは全国のデータ管理者の能力向上を担当する。ANDSはモナッシュ大学、オーストラリア国立大学、オーストラリア連邦科学産業研究機構（CSIRO: Commonwealth Scientific and Industrial Research Organisation）によるコンソーシアムである。

米国では、全米科学財団を含む 22 の連邦政府関係機関を代表するデジタルデータに関する省庁間作業グループ（IWGDD: Interagency Working Group on Digital Data）がデジタルデータの保存およびアクセスに関する戦略計画の策定と普及促進を行っている。最終報告書は本年末に発表される予定であるが、戦略案ではデジタル科学データの保存とアクセスのための全国的調整機関の創設とデジタルデータのアクセス可能性と実用性を最大化するための取り組みを行うことを求めている。教育と研修に関して、IWGDD は 3 つの主要な方針を勧告している。すなわち、第 1 に、教育と研修活動が連邦機関によるすべての科学データへの投資に組み込まれること、第 2 に、全国的調整機関は、連邦機関および教育、研究、技術部門にまたがる教育・研修の調整を推進すること、第 3 に、全国的調整機関は適切な認定と報奨システムを持つキャリアパスとしてデータ科学およびデータ管理を推進することである。

ここ英国では、2007 年春に HEFCE が共有サービスの呼びかけを行い、その結果、英国研究データサービス（UKRDS: UK Research Data Service）の実現可能性を調査するプロジェクトが開始された。このプロジェクトは 2008 年 12 月に報告書が出ることが期待されている。さらに、JISC の助成により 2004 年に設立されたデジタルキュレーションセンターは、その傘下に幅広い活動を行っており、その 1 つに本年末に行われるデータスキル・サマースクールの配信がある。さらに、JISC はデータに関する数多くの研究を委託しており、その中にはデータ科学者の役割に関係するものもある。また、最近、研究情報ネットワーク（Research Information Network）は研究者が如何にデータを作成、公開、共有しているかに関する研究を公開した（Brown and Swan, 2008）。

より広範囲な欧州同盟においては、データ管理とデータ共有は全欧州にわたる基盤的アプローチの一部である。欧州理事会の命令により 2002 年に設立された ESFRI（研究基盤に関する欧州戦略フォーラム）は、汎欧州研究基盤ロードマップを 2006 年に公開し、2008 年 12 月には会議を招集する予定である。

4. 英国におけるデータ科学者の役割とキャリア

4.1 はじめに

データ科学者はコンピュータスキルを実験チームに提供するという主要な役割を持っていると仮定されている[3]。これは真実であるが、これらのスキルやその習得法を検討する前に、よりコンピュータ指向の研究者からは彼らは*必ずしも*必要とされないことに注意するべきである。研究者の中には、学生時代に受けた関連講義から得たデータベースとコンピュータに関する多大なスキルを持ち、Liz Lyon（Lyon, 2007）が 2007 年の報告書で造語した「ネイティブ・データ科学者」に該当する者もいる[4]。Linux認定コースを受けた者もいれば、Java認定資格の取得を続けているものもいる。現在の研究者コミュニティのコンピュータスキルは、まったくスキルを持たない者から何でも知っている者まで、ピンからキリまで存在することも知っておくべきである。

実際、データ科学者のポストについている人は少ないので、通常、研究者はデータスキルを習得する方法を自ら見つけ出さなければならず、普通は、所属の機関や部局の IT サービス部門にアドバイスや支援を求めることになる。これは多くの研究資金を受けている科学分野においても必要だと思われる。例えば、本プロジェクトにおいて我々は 4 つのシステム生物学グループを調査したが、その内、3 グループはデータ科学者のポストを持っており、残りの 1 グループは持っていなかった。このグループでは、研究者はコンピュータ科学者（コンピュータ技官や IT サポートなど、様々な名称で呼ばれている）と連絡を取り、データ処理に関する問題についてアドバイスをもらったり、実用的な解決法を教えてもらったりしていた。このような支援は、自然科学系部局では簡単に得られるが、その他の部局や人的資源にあまり恵まれていない機関では同じようにはいかないと思われる。データ処理やデータ管理に関するスキルが今後益々求められるのは確実であると思われる。

4.2 データ科学者は何をするのか

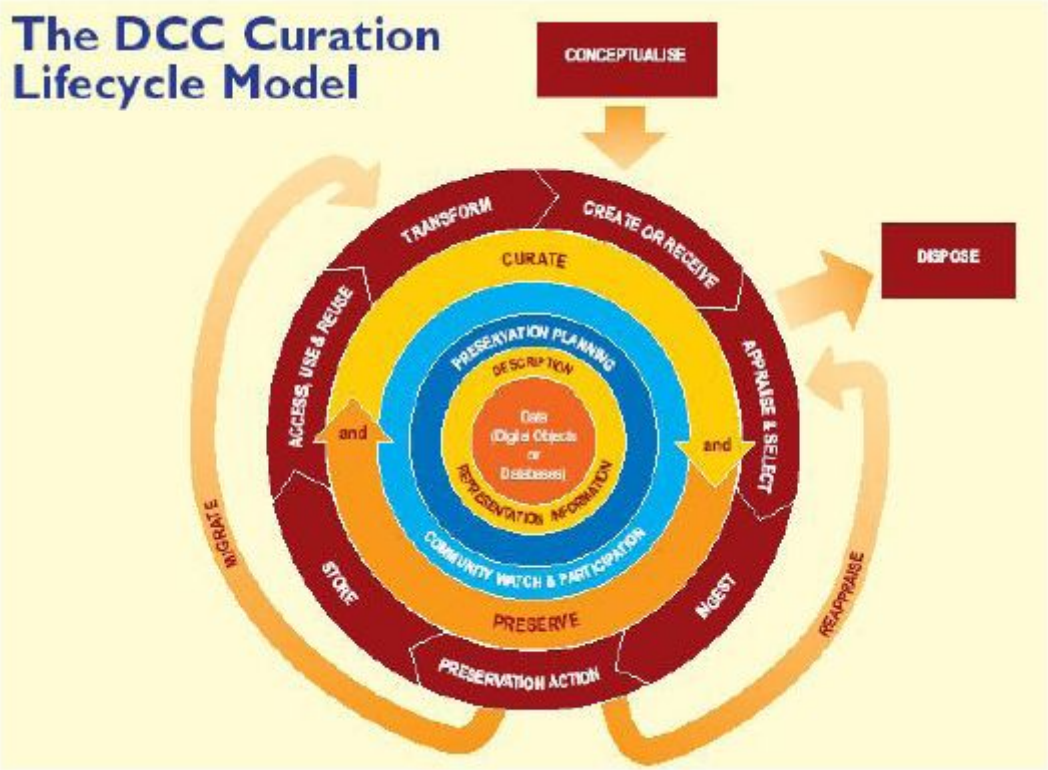
現在データ科学者は少ししか存在せず、その存在も「あいまい」であるが、職務と責任に関する現行制度については、専門的な職務としてのデータ科学の極めて明確な描写が登場していると我々は考えた。

「正式な」データ科学者は何をするのかという点に関しては、データのライフサイクルアプローチを採用し、その中で、NSF で作成され 3.1 節で言及されたデータ科学者の定義に立ち返ることが

[3] ここでは、「実験者」または「実験チーム」という用語を、研究データを生成・作成する作業に関与する人を示すものとして使用している。実験科学以外の学問分野においては、データは別の方法で作成され、別の形態をしている。しかし、ここでは簡単のため、この用語を広義にデータ作成に関与する研究者を示すものとして使用する。

[4] "… 「ネット世代」はデータ共有に関してより解放的である。実際、時間が経てば … 標準的な教育カリキュラムで学んだことにより関連スキルを身に付けた「ネイティブ・データサイエンティスト」が現れると思われる。（Lyon, 2007, p55）

役に立つ。デジタルキュレーションセンターが作成した、データキュレーションのライフサイクルを視覚化した図を以下に示した。その出自を考えれば当然であるが、DCC は、この図の 5 時以降に表現されている、キュレーションと保存作業に重点を置いている。



3.1 節で示した広義の役割定義を使用して、様々な役割がこのサイクルのどこに配置されるかを次のように特定した。

4.2.1 データ科学者

研究グループに所属するデータ科学者は、おおよそ 12 時から 4 時の位置にある作業、すなわち、*概念化*、*作成*、*アクセス*、*利用*、*評価*、*選択*に重点を置いている。また、従事する研究コミュニティの専門分野やタイプ、その研究コミュニティの規範や要求により、さらに別の作業を行うこともある。例えば、我々が詳細に調査した研究分野の 1 つであるシステム生物学では、一般の人がアクセスできる大規模なデータベースが存在し、研究者がデータを投稿すると専門家によるデータセットの保管と保存処理が保障されているので、データ科学者による保存や格納作業に対するニーズは限られている。しかし、他の分野ではこれは当てはまらず、データ科学者がローカルで少なからぬ保存作業をする責任を担うことになるだろう。これらの場合、関連する問題に関する機関としての定まった見解はまだ必ずしも存在しないが、今後の見通しは存在する。

ここで 1 つ覚えておくことは、データ科学者はライフサイクルモデルとは少し異なる用語を使用する傾向があることである。特に、彼らはデータの変更、注釈、変形、派生、その他の操作をキュレ

ーションと呼んでいる。これらの操作によりデータは最も生のステージから各ステージへと渡され、そのあらゆるステージでデータは実験者によりアクセスされ、使用される。普通の言葉で言えば、キュレーションとは、通常は（上のデータライフサイクルモデルに示されているように）保存という用語と結びついた用語であり、長期的にデータの世話をするという目的で行われる活動を意味している。それゆえ、データのライフサイクルのすべてのステージで行われる活動を正確に示すために、本報告書ではできるだけキュレーションという用語を使用することは避け、データの処理と利用にデータ科学者と実験者が関与している活動を示す目的にのみ、この用語を使用する。

データ科学者のスキルは、DCC のライフサイエンスモデルの用語では*概念化*と呼ばれる、データの作成処理を開始する前のステージにおいても極めて重要である。実験科学分野では、データ科学者は実験の計画や設計を支援し、データを収集する最適な方法や収集するのが望ましいデータのタイプについてアドバイスを与え、場合によっては、実験装置やその専用ソフトウェアの使用法をデータ作成者に教え、プロジェクトの研究からデータ関連の最大の利益を引き出す研究プロトコルの策定を支援する。この初期ステージにおいても、データ科学者は研究者と共に作業を行い、プロジェクト提案書にデータ計画を書き、助成団体によるデータに関する要件についてアドバイスを与え、研究が正しいデータ管理作業に準拠した方法で進み、外部や機関の関係者から研究グループに課せられたあらゆる義務を満たすことを保障するように支援する。

実験を行っている間も、データ科学者は 2 種類のデータ出力（例えば、X 線画像を持つ双焦点顕微鏡画像など）の比較や2面登録をする方法を決め、それにより研究プロセスへの洞察を提供して実験者の理解を補足する。ここで最後に指摘しておくことは、グループに参加しているデータ科学者はデータのアクセスや再利用の支援（DCC のライフサイクルではさらに進んだ位置に現れる作業）を頻繁に求められることである。研究グループが生産したものではないデータセットへのアクセスや再利用には、特別なスクリプトの作成やデータ変換といった研究者が必ずしも持っていないスキルが必要になる場合があるので、データ科学者のスキルが必要になるのである。データ科学者はこの時点で役に立つスキルを提供するだけでなく、このような出来事が生じた場合に必要となるスキルを研究グループに教育することもできるだろう。

研究コミュニティの一角を占めているデータセンターに勤務するデータ科学者は、これらすべて（か、その一部）を行っていると思われるが、さらにライフサイクルの 5 時以降に示されている作業にも大きな注意を払っている。通常、データセンターの運営規模が大きくなると、個人はより高度な専門化が必要となる。そのようなセンターでは一人のデータ科学者はデータキュレーションプロセスの一部しか担当しなくなる。このような状況にいるデータ科学者は極めて焦点を絞ったスキルを磨いているが、そのスキルは特定の処理やデータセットに密接に関連しており、移転可能性という意味では限られた価値しか持たない可能性がある。とはいえ、通常、データセンターはデータ科学者に様々な仕事をする機会を与えることができ、場合によっては、データ科学の役割と

平行してオリジナル研究を行う可能性さえ提供する。この多様性は、中期的にはデータ科学者の関心と情熱を持続させる上で重要である。また、データセンター所属のデータ科学者は、研究グループで働く者よりはるかに高度な専門家によるバックアップを受けることができる。通常、データセンターは多様な集合的経験を有しており、大量の専門知識が職員の中に蓄積されている。そして、データ管理者や他の技術者の手を借りることもできる。最後に、ほとんどのデータセンターは運営資金を得るために定期的に助成金申請をする必要があるが、確立したデータセンターは継続的に資金提供を受けられる良い位置にあり、さらに、特定プロジェクトのための追加資金を得る機会を得るためにも良い位置にある。これはデータセンターで働くデータ科学者は概して、他の職場環境に所属する者より、はるかに雇用が保障されていることを意味する。

4.2.2 データ管理者

データ管理者は、我々の定義によれば、コンピュータ科学の専門スキルを持ち、データベース技術の専門家であり、研究チームにより生産されるデータ、あるいは必要とされるデータを適切かつ確実に保管、キュレート、保存する責任を有している。また、データのバックアップやリフレッシュ、必要に応じたフォーマット移行など、システムの管理を行い、どんなシステムを構築する必要があるか、どんな種類のデータを管理する必要があるかについてデータ科学者（や時には直接実験者）と連絡を取っている。通常、データ科学者は実験者とデータ管理者の「仲介役」を勤めることを当然だと考えており、データに関する実験者のニーズや問題点をデータ管理者が理解できる形に翻訳する。

4.3 データ科学者の資格とキャリアパス

現在データ科学を主な仕事にしている多くの人々にとって、これが慎重なキャリア設計や選択的研修の結果でないことは明らかである。概して、この職業に至る道のりは計画的ではなく偶然によるものである。以下のシナリオは本研究に参加した者が繰り返し言及したものである。

- 研究グループ。研究グループにおいて、ある個人がデータ科学機能を担当する主任研究員に指名される。通常、これらの人は情報管理に関する何らかの適性を見せることになり、そのまま続けることを選択し、適当な機会が生じるとそれをメインの仕事とする。ただし、この指名が心底歓迎されない場合もある。キャリアを開始したばかりの研究者は、そのような指名は研究意欲を殺ぐものだと考える傾向があるからである。
- データセンター。インタビューの際、データセンターで働く多くのデータ科学者は、現在の職種を意識的に選んだわけではないと述べた。多くは、公私立の常勤研究ポストに移る前に 1,2 年ほどと思ってデータセンターで働き始めたが、その後も数年間同じ機関に残っている。なぜ彼らは残っているのか。研究者としての背景を持つ者として、彼らは（講義を行う義務なしに）研究コミュニティに参加していることが魅力的であると述べているが、1 つの職場で働くことで助成金獲得の申請をしたり、短期契約で働い

たり、キャリアを守るために様々な国を巡ったりする必要がないという職の安定も好ましいと考えている。

- 主題専門機関。インタビューした機関の管理者は、2 つの問題を抱えていた。1 つは、主題専門知識、技術スキル、対人スキルを併せ持つデータ科学者を採用することが難しいことである。そのため、誰を採用しても必ず追加の研修が必要となる。2 番目の最も重要な問題は、機関に所属するあらゆるレベルの研究者がデータ科学者の必要性を十分に理解しないだけでなく、データ科学者を評価もしないことである。これは、明確なキャリアパスの欠如と共に、データ科学者の不満となり、研究所を去る原因となる。少なくともいくつかの研究所では、データ科学者は仕事をするより自らの役割を説明し正当化することにより多くの時間を費やしていることが明らかになっている。これは、機関の上級管理者やその助成団体が解決しなければならない課題である。

データ科学者が持つスキルにはある程度差はあるが、現在、すべてのデータ科学者は少なくとも各自が参加する専門分野に関しては相当の能力を持っているということができる。例えば、システム生物学やその関連分野であるゲノム学などにおいては、ほとんどのデータ科学者は「その分野の出身」であり、実験室からデータ科学者の職に移動してきた者である。研究チームで働く多くのデータ科学者はその分野の博士号を持っている。実際、データ科学者はグループにおけるデータ関連事項に責任を持っており、その意味では専門家であるが、その多くは今でも自分を実験チームの現役メンバーであると考えている。少数のデータ科学者はその分野以外から現在のポストについたが、そのほとんどは通常コンピュータ科学か情報科学が専門であった。この状況はほとんどの学問分野についても当てはまると思われる。

研究を進める中で、データ科学スキルに関する特別な研修を受けた人に会うことはほとんどなかった。ほとんどは専門家としてのデータ科学スキルを実際の仕事の中で*その場しのぎの方法*で取得していた。とは言うものの、*新規*データ科学者のポストや既存のポストを去るデータ科学者の後任候補者の要件には、必ず何らかの情報学の履修が含まれており、その結果、この役割の形式化と専門家が始まっていると思われる。以下は、システム生物学の研究グループでデータ管理（このグループの用語）を行うポスドク研究員を募集する最近の広告を抜き出したものである。

> *理想の候補者は、物理学またはコンピュータ科学を専攻し、バイオインフォマティクスまたはその関連分野における Ph.D.相当の資格を持ち、データベース技術、ソフトウェア開発、生体分子シミュレーション、顕微鏡関連画像データに関する豊富な経験を有する者である。採用には、優秀な出版業績と生体分子データのデータベース管理に関する折り紙つきの業績が鍵になる。分子動力学データの経験は優遇される。*

情報学が大学院のカリキュラムにおける専門的かつ高度な科目の 1 つになっている分野もある。

この専門化の 2 つの例がバイオインフォマティクスとケモインフォマティクスである。両者は、英国の多くの大学で修士レベルの科目として提供されている。しかし、情報学の研修が必ずしもデータ科学者に必要でないように、これも万能薬ではない。我々が話したデータ科学者は、少なくとも修士レベルの情報学教育は、その分野におけるデータ科学者の役割に必要なレベルの研究知識を提供していないという意見であった。大学院で情報学コースを教えているある研究機関の所長は、このコースはデータ科学の役割に必要な知識を卒業生に与えるものではないと述べた。仕事についてからさらに個人的に開発したり学習したりすることが必要である。

本研究のオンライン調査に参加した大多数のデータ科学者は修士号を持っており、およそ 1/3 はもう 1 つ別の関連する大学院レベルの資格を持っていた。必要とされるスキルの組み合わせについては、各自が働く主題分野に関する学位や高度な資格を持つことがデータ科学者にとって不可欠であるか否かについて、データ科学者の間で全体的な合意がなされていないことを、本研究は明らかにした。

現在データ科学者の役割についている者は、高度なコンピュータスキルを獲得した主題分野の専門家か、仕事の中で主題に関する知識を吸収したコンピュータ科学または情報科学の専門家か、のいずれかであろう。（少なくとも実験科学の場合）前者が一般的であるが、圧倒的多数がそうであるわけではなく、時間をかけて必要とされる主題知識を吸収することによりコンピュータ科学者がデータ科学者の任務を果たすことは可能であり、実際に果たしている。我々はある研究グループの一人のデータ科学者に出会った。彼はコンピュータ科学者であり（現在の職を見つける前に）ある主題関連の研究グループに参加し、その分野に固有の最新の概念や方法論を理解するまで 2 週間かかったと述べた。

使い物になるデータ科学者になるためには関連分野における何らかの研究経験が必要であるか? 研究者としての背景を持つデータ科学者の答えは、はっきりと「イエス」であった。彼らは、従事する主題分野を詳細に理解していないと現在の役割をこなすことは非常に難しいだろうと主張する。この意見は、特定の主題分野に関係する情報の具体的な特徴を理解することの必要性に基づいているが、同時に、主題分野に関する深い知識を持つことがデータ科学者と研究グループの他のメンバーとのコミュニケーションを容易にするということにも基づいている。また、データ科学者になった（自分の名前を冠した業績や出版物を持つ）研究者は同僚から尊重される傾向があること、この同格性は、それがない場合に比べて、データ科学の処理をよりスムーズかつ効率的に行うことができるようにすることが報告されている。

いろいろな意味でこれは理想的なシナリオであり、研究者がデータ科学に集中するというキャリア上の迂回路を選択するものである。ただし、1 つ問題があるように思われる。通常、研究者はそのような迂回路を望まず、たとえそうしたとしても、いつかは常勤研究者に戻りたいと考える者がいる

と思われることである。これはデータ科学の価値を認識していないからではなく、研究により提起される課題に再度挑戦したいと考えるからである。幸運な者は、仕事としてデータ科学と研究の両者を行う機会を得ている。

一方、仕事を効率的にこなすために主題専門家になる必要はないと主張するデータ科学者も存在する。実際、秘密研究の処理、データ記述とメタデータ、ソフトウェア、著作権と知的所有権、データ保存など、基本的なデータ科学スキルには汎用的なものもあるからである。これはおそらくそうであろうが、一番重要な問題は、データ科学者と研究者との効率的なコミュニケーションに関するものである。例えば、芸術分野の教育を受けた者が天文学データを扱うのは大変であり、必要とされる洞察力やスキルを習得するために必要な時間は研究グループのスケジュールや予算にとって有害であろうことは誰もが認めることである。データ科学者は関連の主題分野における大学院レベルの教育を受けるべきだというのが妥協点であると思われる。実際、これは修士レベルの図書館学コースの科目としてデジタルキュレーションを提供するライブラリスクールが増加している点と一致する。

実用的な観点から言えば、優秀なデータ科学者への需要が増すにつれ、できるだけ幅広く網を打つ必要が出てくる。主題知識は重要であるが、技術スキルや対人スキルも同じく重要である。主題知識の不足は効率的なコミュニケーション能力を持つ人なら埋め合わせることができる。研究グループを味方に付け、効率的な協力関係を構築するために必要な個人的資質を持ったデータ科学者は極めて生産的になる可能性がある。

また、技術やコンピュータに対する適性の問題も検討しなければならない。データ科学者は、データフォーマットやデジタル化、データ保存、データアクセス、セキュリティなどの問題について詳細に理解していることが求められる。通常、その役割は実用的なコンピュータスキルだけでなく、例えば、機関のコンピュータセンターの職員と対等に話し合うことができるだけの技術用語に関する知識が必要とされる。本研究は、ほぼ半数のデータ科学者は、その職務をする人にとって技術またはコンピュータに関する経歴を持つことが不可欠であると考えていることを示した。

また、現在データ科学者として働く人々へのオンライン調査では、データ科学者として成功するには対人スキルが技術スキルより重要か否かについて、データ科学コミュニティの意見は2つに分かれた。通常、人々の意見は、各自の経験と各自の長所が技術スキルと対人スキルを両端にもつスペクトルのどちら側に位置しているかに基づいている。両スキル共に優れている人はあまり見かけない。我々は、主にコンピュータと情報技術を背景に持つ人が、所属する専門機関の主題分野を十分に習得し、有能なデータ科学者と考えられるような存在になっている実際例にいくつか遭遇した。上級ポストにいる者の中には、技術的な背景を持っている人に主題分野の基本を教える方がその逆よりおそらく良いだろうという意見を持つ者がいた。そのマイナス面は、もちろ

ん、そのような知識を習得するにおそらく何年もの長い時間がかかる可能性があることである。研究分野の中には、求人数よりはるかに多くの高学歴な人が存在する分野もある。通常、そのような状況では、雇用者は博士レベルの主題知識に加えて、修士レベルのコンピュータまたは情報技術のスキルを提供する候補者を見つけることが可能である。

4.4 データ科学者を抱えるポスト

研究データが生産される場所であれば、どんな立場であれデータ科学者の仕事があると考えることができる。以下に、主な仕事の種類を示した。本研究の一環としてデータ科学者に、専門家としての達成感という観点からみてどんな種類の職種が魅力的であるかを尋ねた。もちろん、その結果は個人的な嗜好や経歴から生まれるものであるが、現在データ科学者が考える魅力的な組織に関する有用な概要を提供する。最も人気のあった上位 3 つの選択肢は、次の通りである。

- 主題専門研究機関のデータ科学者
- 研究会議に勤務するデータ科学者
- 研究プロジェクトチームに参加するデータ科学者

最も人気のなかった下位 3 つの選択肢は次の通りである。

- コンピュータセンターに所属するデータ科学者
- データサポートサービスに勤務するデータ科学者
- 大規模データセンターのデータ科学者

中位を占めたものは、図書館と小規模データセンターに所属するデータ科学者であった。

4.5 雇用保障

調査したデータ科学者は任期の保障に関して主に 2 つのカテゴリに分けられる。第 1 は、大学、研究センター、データセンターにおける終身雇用のポストに従事する者であり、第 2 は、短期間の研究助成金により雇用されている者である。

4.5.1 大学と研究センターにおける終身雇用のデータ科学職

データ科学者の中には、英国の高等教育機関において完全終身雇用のポストを確保している者がいた。その立場は様々である。そのようなポストに与える具体的な職級を持っている大学もあり、その場合、教員職と技術職のいずれかの職級に該当する。その一例は、インペリアル・カレッジ・ロンドンの*専門サービス*職である。教員に順ずる職（英国の大学で普通に見られる「教員関連」職と必ずしも同じではないのでこの用語を使用する）にある者が教員職や教員関連職に異動することは難しいと思われる。学術研究の「価値」を測定するために通常使用されている出版物を作成していないからである。他の大学には終身雇用のデータ科学者が教育職である場合もある。ただし、証言によれば現時点ではその数が非常に限られていることが示されている。データ関連

のプロジェクトがより一般的になり、データ科学が大学において重要になれば、この問題に対する大学の考えもいくらか変わることが期待されるかもしれない。そうなれば、高度なスキルを持った専門的なデータ科学者の新規採用と雇用確保に対するニーズは否応なくより切迫したものになるだろう。

4.5.2 短期契約のデータ科学者

多くのデータ科学者は短期契約で雇用されている。ただし、その任期は必ずしも 1 回限りのものではない。雇用が更新される場合や、例えば、大学で運営される天文学データセンターの場合のように、プロジェクトを運営している機関の中で次期助成金申請に含まれているポストに配置換えされる場合がある。

短期ポストが完全に相応しい場合もある。行われるべき個別の研究があり、それが実行され、プロジェクト実施期間中はデータが適切に管理・キュレートされ、その後、将来の再利用のためにどこか別の安全な場所にデータが保管される場合である。GenBank のような適当な公的データバンクに登録されているデータを持つ学術研究の他の多くの領域において、そのような取り決めは満足すべきモデルであろう。その他の場合においては、天文学の現状がその良い例であるが、非常に高齢のデータ科学者が何度も契約を更新している。彼らはデータ科学の専門家であり、何年もかかって習得した極めて高度なスキルを持つ人々である。彼らはその研究グループの運営には不可欠であり、グループ内で非常に高く評価されているが、ほとんどの場合、大学における「正式な」身分を持っておらず、昇進コースも定まっていない。

4.6 研究コミュニティへのデータ科学スキルの提供

研究データスキルの価値が研究コミュニティにおいて明らかになるにつれ、また、助成団体が研究プロセスに対する投資の見返りをより多く求めるようになるにつれ、データ科学者の役割の重要性は増すことになる。現時点で、短中期的に増大するニーズを満たすだけの数の適切なスキルと経験を持つデータ科学者が現れるかは疑わしい。現在は、需要が供給を上回っている。2 つのグループからは、求人広告を出したが適当な人を雇うことができなかったという話を聞いた。その理由は元来この職が魅力的なものではないからではない。少なくとも、調査に参加したデータ科学者の半数は自分のキャリアが専門家として報われるものであると考えており、およそ 1/4 は自らの職の高い自律性に満足していた。ただし、英国のデータ科学者が現在働いているキャリア構造については基本的な問題が数多く存在するように思われる。オンライン調査の参加者により指摘された主要な問題は以下の通りである。

- データ科学者の 2/3 はデータ科学におけるキャリアを継続するために何らかの正式な研修が必要とされるのは当然であるという考えに反対である。

- 多くのデータ科学者は、データ科学におけるキャリアを継続したいと考えている者に正式な研修が提供されているとは考えていない。

- データ科学者の半数以上が、データ科学の役割が研究コミュニティにおいて理解も尊敬もされていないと感じている。

総合的に見て、研究チームは次の 3 つの活動を重点的に行っているようである: 助成金申請書を書く; 研究を行う; 論文を書き出版する。データ科学の問題は優先リストのはるか下の方にあり、場合によっては無視される。その自然な結果として、データ科学者により担われている役割は必ずしも高く評価されず、上に挙げた 3 つの重点活動のすべてにおいてデータ科学者が益々重要な役割を果たしているという事実にもかかわらず、本研究で調査した多くの事例においてデータ科学者の役割はプラスに評価されずに費用負担だと考えられていることが報告されている。データ科学の重要性が多くの研究者に理解されていないこと、それに付随して、研究チームで働くデータ科学者の役割が専門家として十分に尊敬されていないことは、その契約条項に反映される可能性がある。

オンライン調査の回答者により挙げられたデータ科学のキャリアを選択または継続する気をなくさせるその他の要因は次の通りである。

- 英国には常勤のデータ科学職が数多く存在するという考えに同意する現役データ科学者はほとんどいない。
- 1/4 以上のデータ科学者がデータ科学職は短期間のものが多いと考えている(半数はわからないと述べている)。

- データ科学職は教育職というより技術職だと考えられる傾向にあり、全体として、データ科学職に公正な報酬が与えられていると考えているデータ科学者は 10%に過ぎない。

キャリアアップの問題に関しては、

- 2/3 のデータ科学者が、データ科学におけるキャリアを継続したいと考えている人に対する明確なキャリアパスが存在しないと考えている。

現時点では正式なキャリア構造は存在せず、もちろん、学術研究者や技術者のキャリアとは比較できない。それどころか、大学でデータ科学者として働いている人々は、従来の意味におけるキ

ャリアアップはほとんど期待できない特別な仕事をしている。もちろん、新しいスキルを習得することにより現在の職務の質を高める事はできるが、現在の職場を離れる事を決心する際、その異動は横滑り、すなわち、別の組織の同様なポストへの異動になる傾向にある。職階や賃金レベルという意味におけるキャリアアップの範囲は非常に限られているからである。実際、データ科学者は、非常に専門的なスキルを磨くことは現在の職場とさらに密接に結びつくことになると報告している。そのようなスキルは他に応用が利かないことを知っているからである。この状況は、今後、そのようなスキルと経験がより広く評価され、必要とされるようになれば変化するかもしれない。

データセンターに勤務している場合は幾分状況が異なる。そこでは、職員は公共サービス職であり、キャリアアップの可能性が存在する。もっとも実際は職階が比較的フラットであるので、職階が上がるには何年もかかる可能性がある。少数の優秀な者はこの過程を短縮することができるが、通常、若い新入社員は自らのキャリアに対する要望と組織的制約による現実とのギャップを感じて長く留まることはない傾向にある。

労働市場においてデータ科学の経験を持つ研究者が「付加価値を伴う」競争力を持つ可能性があることを示す証拠が存在する。最近、英国のほとんどの研究会議は、助成を行う研究チームがデータ計画とデータ科学一般に十分な注意を払うことを要求するようになってきた。効率的なデータ科学の実務へのニーズが徐々に研究コミュニティに浸透するようになるにつれ、データ科学スキルはさらに評価が高くなるだろう。現時点ではそのような供給が不足しているので特にそうである。しかし、既に注意したように、現在データ科学の問題を担当している研究者の中には元々のキャリア（研究を行うこと）に常勤職として戻りたいと考えている者がいる。この傾向は、研究者としての経歴を持つデータ科学者の予備軍を限定することにつながるだろう。しかし、研究の魅力に惹かれなくなった者や比較的安定したデータ科学職を魅力的だと考える者も存在する。この文脈において、例えば、育児休暇を終えて仕事に（できればパートタイムで）戻りたいと考えている研究者にとってデータ科学のポストは魅力的であることが報告されている。

5. 研修の提供

5.1 はじめに

研修コースに参加させる気にさせるために重要だと考える研修のやり方に関しては、データ科学者は極めて実用主義的である。オンライン調査によれば、彼らは特に次のような研修を高く評価する。

- 実務家による実体験の講演
- 現在の仕事に直接関係する実用的スキルの習得
- 参加型で実用的な練習課題の提供

また、データ科学者は現在持っているスキルが数年後には役に立たなくなるかもしれないことを理解している。この分野における変化は非常に早いので、データ科学者がこの分野の進展に遅れずについていくには、継続的なスキルアッププログラムが必要になる。

データ科学者は研修コースへの参加を正式に評価することはそれほど望んでいない。現役のデータ科学者で、研修コースを終了した正式な証明書や認定を受けることが重要だと考える者はおよそ 1/4 にすぎない。この各自の仕事に適用できる実用的な研修への嗜好の偏りは、データ科学者の「仕事を通じて」学習するという伝統を反映している。以下では彼らの嗜好についてより詳細に検討する。

本章では、データ科学者向けだけでなく、専門家としてのデータキュレーションスキルを研究コミュニティに提供するその他の種類の研修に関する概要も提示する。我々が RIN、JISC、NERC のために行った調査（Brown and Swan, 2007）とオーストラリアの大学におけるデータ作業に関する最近の研究（Henty et al, 2008）は共に、研究者は全体としてデータの共有を好み、各自のデータをより良く管理することの重要性を理解し、一般に良いデータスキルを持つためにさらに勉強したいと考えているが、そうするための支援を必要としている証拠を提供した。これらは研究者が何かのついでに容易に実行することができることではなく、これを研究者は理解している。さらに、必ずしもデータ科学者が行う必要がないデータ関連のスキルも存在する。データの長期的な保管・保存は図書館の役割だと考えたほうが正しいと思われる。これは図書館の活動範囲の一部である機関リポジトリ（場合によっては、独立した機関データリポジトリ）が本領を発揮する領域だからである。もちろん、これは関連するデータスキルが図書館コミュニティに備わっていなければならないことを意味する。これらの問題をこの章で検討する。

5.2 データ科学者向けの OJT スキル開発

データ科学者は永久にやっていけるだけのすべてのスキルを完全に身につけた状態で職に就

いたわけではない。彼らは常に最新の技術や情報を習得する必要がある。本研究の調査に参加した大多数のデータ科学者は仕事を「実務の中で」習得したと述べている。実際、我々の調査は、回答者の 2/3 が独学であったことを示している。データ科学者は実務の中で様々な方法によりスキルを習得することができる。これらのスキルはまさに、日々の業務の中で実験者から与えられる新たな課題に対応するために、実験機器メーカーにより実施される限定的な研修コースなど、たまに行われる短期コースを通じて形成されたものであると思われる。専門分野に特化した短期コースという考えはデータ科学コミュニティにおいて大きな関心が持たれている。これについては 5.4 節でさらに検討する。ある種の組織、特にデータセンターでは、データ科学者はアドバスが必要な場合に、より経験をつんだ同僚の専門家としての意見を聞くことができる。

データ科学は、技術スキル、対人スキル、主題知識という一般的ではないが価値の高い組み合わせの資質を併せ持つ必要があることを考えると、既存のデータ管理者コミュニティの大部分が、このどちらかといえば非公式な実務上のルートでスキルを取得したことは興味深いことである。特に重要な理由として次の 2 点を特定しているオンライン調査の回答がその理由をある程度示している。

- 適切な研修機会の欠如（および、研修に関する知識の欠如）

機関規模であれ、小さな研究グループ規模であれ、各組織は異なるニーズと優先順位を持っていることがデータ科学の特徴である。通常、データ科学者に指名された者は、自らのニーズは極めて特別またはユニークなものなので、現在利用できる研修コースはこれらのニーズを十分に満たすものではないと考える。実際、我々が連絡を取ったほぼ半数のデータ科学者は、自分が必要とするレベルの研修機会がないと述べた。一方で、ほとんどのデータ科学者はデータ科学やデジタルキュレーションに関する研修コースやスキルアップコースに参加するために時間を費やす準備があるとも述べている。調査した 1/3 強のデータ科学者は 5 日を越えるコースに参加する用意があると述べ、調査の過程で話した者の中には、この夏に欧州で何日にもわたって開催される様々な研究データスキルアップイベントに参加登録をしている者もいた。このように、長期コースという考えを歓迎する者もいるが、全体的な嗜好は短期集中コースであり、これはほとんどのデータ科学者に評判が良いと思われる概念である。

- 研修イベントのコストと開催場所

研修に対して不利に作用するもっとも一般的な 2 つの要因は、時間の不足と資金の不足である。たとえ適当な研修コースがあっても、移動にかかる時間やお金は、およそ 1/3 のデータ管理者にとって大きな障害である。比率は低いが、組織的な支援がないことも外部の研修コースに参加する際の障害だと考えられる。

5.3 正式な大学院教育

英国の大学で徐々に提供されるようになってきた専門的情報学の教育に加えて、他の専門分野の修士課程がデータ科学の要素を含むようになってきた。大量のデジタルデータを生成する専門分野の修士課程のほとんどは、現在、研究教育実習の一部として、データ処理、データ管理、データキュレーションを含んでいる。さらに、博士課程のプログラムは、今後常に、情報の収集、記録、操作、格納などに関するその分野固有の実務に必要となる水準の教育を含むことになるだろう。これらすべては、データスキルが益々重要になるにつれ、継続・増加することが期待できる。

我々は、データ科学に関する正式な大学院教育を受ける可能性のある 2 つの客層を区別する。1 つは、データ科学者になりたいと考える者であり、もう 1 つは、研究者としての道を離れ、仕事としてデータ科学の世界に入る気はないが、データ関連のスキルアップをすることにより利益を得ることができる大多数の研究者である。

5.3.1 データ科学者向けの教育

既に述べたように、調査に回答したほとんどのデータ科学者は計算機科学や情報学といったデータ科学に関係する分野の大学院レベルの学位を持っている(もっとも、専門分野の教育を受けただけで、データ専門家になるための勉強は実務の中で行ったというデータ科学者も数多くいた)。もっとも関係のある大学院レベルの資格は、情報学の修士号である。ますます多くの大学がこの学位を提供するようになっている。一般的な情報学の学位が計算機科学や情報科学の学科で与えられている場合もある[5]が、今日では多くの場合、そのようなコースは専門分野に特化したものになっており、そのもっとも一般的な例は、バイオインフォマティクスとケモインフォマティクスである。今後、多くの分野がデータ中心的な研究分野に移行するにつれ、他にも専門分野に特化した情報学コースの数が増加するだろう。

情報学の教育におけるそのような専門分野指向のアプローチに関して重要なことは、データの操作や利用には専門分野固有の強い特徴を持っていることである。例えば、分子生物学の研究で生み出されるデータセットは各々非常に異なったものであり、データセット間の統合(マッシュアップ)や相互問い合わせはこの分野の優秀なデータ科学者が行っても非常に難しい作業である。これらのデータセットと、人工知能や天文学、人類学、考古学、地域研究などのデータセットとの類似点はまったくなく、特定の専門分野のデータセットを操作するのに使用される技法と他のデータセットで使用される技法では、もっとも基本的な部分が似ているだけだと思われる。それゆえ、研究が益々データ中心的になり、データ処理スキルが研究者の道具の 1 つになるにつれ、専門分野に特化したコースが必要とされるようになる。

[5] 例えば、ストラスクライド大学を参照: http://www.gsi.strath.ac.uk/

5.3.2 研究者向けの教育

大学院レベルのすべての研究者にデータスキルの教育を提供する必要があることは明らかである。この種の教育は3つの潜在的利益を持っている。第1に、研究者がデータのライフサイクルの重要性と、研究データの生成、処理、キュレート、保管、保存を成功させるために果たさなければならない役割を理解するのを助ける。第 2 に、研究者の中からデータ科学者になろうとする者が出る刺激を与える可能性がある。第3に、この種の教育は、仕事を開始するために必要な基礎知識を新人データ科学者に授けることになる。

データ科学やデータ管理に関する具体的なコースは一般的にはまだ見つかっていないが、そのようなコースが登場し始めている。その 2 つの例が、キングス・カレッジ・ロンドンの人文学コンピューティングセンターにより提供されている新しい修士課程[6]とグラスゴー大学にある人文学高度技術・情報研究所で教えられている情報管理・保存学理学修士課程[7]である。

学位を提供しないコースも利用できるようになっている。既に、UK Data Archive はデータ科学とデータ管理の主要分野を扱う一連の研修・支援資料を作成している。8 つのモジュールが利用許諾書の作成、匿名化技法、データ記述とメタデータ、データフォーマットとソフトウェア、著作権と知的所有権、データの保管・バックアップとセキュリティ、データのデジタル化とアクセスの提供といった主題を扱っている。資料は、社会科学、経済学、人文学の研究データを生産、処理する研究者向けに企画されているが、そのコンセプトは一般にすべての研究データに適用可能である。UKDA は既に数多くの研修イベントを実施しており、今後もさらに計画している。しかし、最終的には、資料が様々な主題分野の講師に採用され、英国中の研究者に行き渡ると思われる。

デジタルキュレーションセンター(DCC)も大学院教育の提供において重要な役割を演じている。現在、DCC はデータライフサイエンスモデルに焦点をあわせた一連のモジュールを完成させている。デジタルキュレーション 101 と呼ばれることになるコースは、主に研究室の研究者と情報専門家向けに企画されたものであるが、このモジュールはあらゆる専門分野の研究者に合わせて適合させることができるものと思われる。

最後に、最近、英国と米国における現状のベストプラクティスを利用してデータキュレーションの研修・教育を提供することを議論するために作業グループが召集された。国際データキュレーション教育アクション(IDEA: International Data curation Education Action)作業グループとして知られているこのグループは、適当な研修・教育の開発・提供を前進させるための最良の方法およびデジタルキュレーションをそれ自体専門的職業として普及する方法を検討している。IDEA の目標は、参加者を拡大し、会議やメンバー間の持続的な協力を通じてその作業を継続することであ

6 http://kcl.ac.uk/iss/cerch/teaching/
7 http://www.hatii.arts.gla.ac.uk/imp/index.htm

る。

5.4 継続的な専門能力開発（CPD）

既にデータ科学職にいる者には特定のトピックに関する専門的な研修が必要である。より一般的なアプローチはどうしても価値が限られてしまうだろうが、各専門分野で共通の課題をデータ科学者が克服するのを支援することは非常に役立つと思われる。これらの人々が一堂に会することを容易にすることが主要な任務の 1 つであろう。同僚から学ぶことができることはたくさんあるからである。例えば、データ統合は大きな挑戦であり、かつ、非常に重要なことである。現在、異なる起源のデータセットの統合の学習は、仕事の中でその場しのぎの方法や他のデータ科学者の経験や実務上の注意点を交換することにより何とか行われている。ワークショップ形式のイベントを提供することにより、そのような交流や情報交換を定式化することは、高度なスキルをこのコミュニティに拡大するのに貢献すると思われる。

データ科学コミュニティには、データに関する課題は急速に変化するので、フォローアップのない初期の集中的な研修より、定期的なスキルアップ研修の方がはるかに効率的だという意見が多い。この目的を果たすために、既に、様々な組織がデータ関連の話題を扱う研修コースやワークショップを開催している。いくつかの例を次に示す。

- デジタルキュレーションセンターは、最近、英国の様々な専門分野におけるデータ科学者とデータ管理者により構成されているグループ、研究データ管理フォーラム（Research Data Management Forum）[8]の第 1 回会議を開催した。このフォーラムは、経験とベストプラクティスを交換する機会をデータ専門家に提供する。
- 生物多様性および生態情報学のための国際共同サイバーインフラストラクチャである汎アメリカ高等研究所（Pan-American Advanced Studies Institute in Cyberinfrastructure for International Collaborative Biodiversity and Ecological Informatics）[9]
- 生命科学におけるデータ統合に関するEvryワークショップ[10]
- ERASysBioイニシアティブのデータ管理サマースクール[11]（BBSRCにより一部助成された）

英国においては、データ科学コミュニティは専門能力開発のニーズに対処するCPDプログラムという概念を好意的に見ている。データ科学は確かに急速に変化する分野であり、関心のある話題に関する正式な教育の機会に加えて、一堂に集まって経験や実務を議論する機会は、評判の良い考え方である。

このような専門的なCPD研修を提供するのに相応しい機関は数多く存在する。デジタルキュレー

8 http://www.dcc.ac.uk/events/data-forum-2008/
9 http://ciara.fiu.edu/eco/
10 http://dils2008.lri.fr/
11 http://www.erasysbio.net/

ションセンターはその1つある。コンピュータスキルの腕を磨く必要がある場合は、国立eサイエンスセンターや英国コンピュータ学会が CPD 研修を提供する候補機関である。データ科学者が好む CPD 研修のタイプは、最新の話題に対象を絞った集中的な短期コースである。我々が話をした人々は、そのような研修を提供する機関は、特定の専門分野におけるデータ問題(データセットの統合や特別なソフトウェアの操作法など)の専門家をそのような研修の講師にすることを検討すべきだと提案した。2,3 日から 1 週間のコースがもっとも好まれる。

医療データの管理を専門とする人々へのCPDの提供は既に確立しており、他の分野でデータ科学を専門化する際のモデルとなるだろう。臨床データ管理協会(Association for Clinical Data Management)[12]は、(NVQ資格を含む)様々な教育・研修プログラムの提供に加えて、ネットワーク上で会議やSIG、技術会議を行う機会も提供している。

このような研修の提供と共に、研修への長期的関与が必要とされることが機関や助成団体レベルで認識される必要がある。データ科学は急速に変化するので、最先端に居続けるには定期的かつ永続的なスキルアップが必要になる。

5.5 学部教育カリキュラム

今後、データ処理スキルが益々研究者としての基本的スキルの一部となっていくことはまちがいない。「ネイティブ・データ科学者」はすべての学問分野で当たり前になるだろう。既に多くの学部教育課程でデータ関連トピックの講義や単位が組み込まれている。これらはマイクロソフト Excel の操作法や統計学の基本など極めて基本的なものだと思われるが、全体像について考え始めることに早すぎるということはけっしてないという意見に、データ科学に携わるものは皆同意している。実際、我々が話したデータ科学者は、たとえ基本的なものでもいいからすべての研究者がデータ管理スキルを持つようになれば、データ科学者が日々直面している問題のほとんどはなくなり、もっと困難で専門的な仕事をすることができるようになるだろうという意見であった。

この問題には 2 つの意見がある。1 つは、学部教育カリキュラムは、必須の科目をさらに追加しなくても既に十分満杯であると考える者がある。この意見を持つ者は、我々が話をした中では少数派であったが、その指摘は正しいものである。少なくとも、当該分野の学位を与えるのに値するだけの専門知識を教えるだけで時間的にはいっぱいであり、研究関連のスキルは重要だとは思うが大学院の教育コースや博士課程のプログラムの一部とするべきだという考えを持つ専門分野が存在する。もう 1 つの意見は、データスキルは基本的な統計や研究法、実験結果の記録法などと同じように学部教育の基本の 1 つとしてみなすべきだというものである。最終的には後者の考えが広がるだろうと思われる。一般的に教えることができる基本的なデータスキルや原理が存在する。関係データベース、XML、キュレーションの原理、文書作成、作業管理などである。各専門

[12] http://www.acdm.org.uk/

分野の研究プロジェクトでこのようなスキルが必要になれば、これらが大学教育カリキュラムに組み込まれる場所が見つかるかもしれない。

5.6 図書館の役割

図書館と情報科学コミュニティは、データ科学分野で果たすべき重要な役割を持っているはずである。特に、データ問題の周知と理解および良いデータ科学と良いキュレーションの重要性を広めることである。図書館員が持つべき一般的なデータ処理およびデータ管理スキルが存在し、これらは機関における基本的な研究スキル研修で教えることができる。要するに、データ科学の基本は教えることができ、主題専門知識は時間をかけて習得することができる。図書館がこの分野で果たすことができる別の役割が存在する。図書館コミュニティがこの分野の発展に影響を及ぼす可能性がある 3 つの方法を提案する。

- データに対する意識をより高めるように研究者を教育する
- データを保管・保存する役割を引き受ける
- データ図書館員の養成と供給

5.6.1 データに対する意識をより高めるように研究者を教育する

これには、重点を変える必要があるだろう。例えば、英国のライブラリスクールでは既に学部と共同して、一般的な研究関連スキル（情報リテラシーなど）や時には非常に専門的なプログラム（ケモインフォマティクスなど）を教えているが、これは一般にライブラリスクールはデータより情報に関連するものだと考えているからである。さらに、研究プログラムに参加することは、明らかに学部教育に参加することより実現が難しい。例えば、通常、図書館は学部学生に情報リテラシープログラムを提供しているが、これらを研究プログラムに持ち込もうと考えることは一般的でなく、たとえするにしても、それを強制することはほとんど無理である。そうではあるが、データの氾濫は事態を変えるかもしれない。最新のオーストラリアの研究（Henty et al, 2009）は、研究コミュニティが良いデータ実務に興味を持ち、もっと知りたいと望んでいることを明確に示した。この分野に関する我々の研究もこれを裏付けている（Brown and Swan, 2007）。また、本研究で話を聞いた図書館員からも、データ管理に関する実践的な支援を研究者から求められることが多くなっているという話を聞いた。図書館はデータに対する意識を高めるようギアを入れるべきである。そのような需要が既に高まっているからである。

5.6.2 データケアの役割を引き受ける

データの保管・保存能力に対するニーズの高まりは、図書館が研究に対する自らの立ち位置を変える戦略的機会を提供する。図書館コミュニティにはこの話題（例えば、Steinhart et al, 2008 を参照）や図書館が最適なeサイエンス計画を提供できる方法（Martinez, 2007; Carlson, 2006）に関する数多くの議論や計画が存在する。多くの図書館は機関リポジトリの構築という挑戦に応

じたが、その自然な拡張はデータの領域に及ぶ。多くの図書館員は、機関に代わってデータを管理する役割を果たすよう図書館の立ち位置を変えようとしている。我々がインタビューしたデータ科学者は、図書館は実際にデータの保管と保存に責任を持つべきだと考えていた。データ科学者はこれにより研究者と共に別の（分野固有の）データ課題に取り組む時間ができると考えている。しかし、保管・保存作業と全体として表現されるデータ科学には違いがある。図書館管理者は、データ科学という意味で主題図書館員やリエゾン図書館員が提供できるものは限られると警告している。しかし、図書館職員とデータ科学者が互いに学習できる範囲は広く、潜在的な相乗効果を有効に利用することで学術コミュニティが全体として新しい作業方法に移行すれば、これが当然のことになることが期待される。データセットをハーベストしてキュレートする大学図書館の潜在的な役割の評価は本研究の範囲を超えるものであり、DISC-UK DataShareプロジェクト[13]で検討されている。

5.6.3 データ図書館員の養成と供給

図書館学の教育者は、データキュレーションという図書館の潜在的役割をデータ図書館員が果たすという潜在的需要を満たす適切なスキルを持つ者を養成し、供給するという重要な任務を持っている。しかし、将来のデータ図書館員が必要するスキルを教えているライブラリスクール（図書館・情報学専門大学院）は、現在ほとんど存在しない。

例えば、米国で情報・図書館学を教えている 55 の機関のうち、その課程に何らかのデジタルキュレーションの内容を含んでいる機関はほんの一握りである。ノースカロライナ大学チャペルヒル校ではデジタルキュレーションのカリキュラム[14]を作成するためにかなり多くの努力を注いでいる。このチームは、この件に関して議論するための会議を組織している[15]。イリノイ大学ウルバナシャンペーン校の科学・学術情報学研究センター（Center for Informatics Research in Science and Scholarship）[16]のチームも、図書館員がデータキュレーションの役割を果たす最適な方法を考える作業を行っている。英国では、シェフィールドのライブラリスクールがケモインフォマティクスを教える短期コースの教育に参加しているが、一般に、ほとんどのライブラリスクールはデジタルデータ管理を図書館学修士課程の 1 科目として含んでいるが、この分野における特別な選択肢は提供していない。英国にはデータ図書館員と呼ばれる者が数人いるが、その数は現在およそ 5 人だと考えられる。

数多くのアプローチがあるにもかかわらず、ライブラリスクールが未だに十分なデジタルキュレーションスキルを持った図書館員を生み出していないのには数多くの理由があると思われる。

13 http://www.disc-uk.org/datashare.html
14 http://ils.unc.edu/digccurr/aboutI.html
15 http://www.ils.unc.edu/digccurr2009/
16 http://cirss.lis.uiuc.edu/index.html

- 現在、データ図書館員には明確なキャリアパスが存在しない。授業料や生活費を払わなければならない大学院生にとって、通常のライブラリスクールのカリキュラムからはるかに離れることには盲目的な信仰が必要になるだろう。多くの潜在的学生は、この分野についてほとんど何も知らず、このコースがどのようなものかも卒業後に関連する仕事を見つけることができるのかもわからない。この先行きの不透明性からそのようなコースは危険すぎると思われることになる。デジタルキュレーションカリキュラムの基礎を構築する多くの価値ある作業が行われてきたが、これが規範的カリキュラムとして抽出されるまでにはまだ道半ばである。これは、なぜ先導的なライブラリスクールでさえ一般的な図書館学修士（MLS）課程においてデータキュレーション関係を「目立たないように」しておかなければならないかを説明する。MLS 課程に単にデジタルキュレーションの科目を入れておけば、学生は両面作戦を採ることができる。ライブラリスクールは、「図書館情報技術」におけるキャリア（この数はわかっている）の見込みを含めて、想定される学生にデジタル関係のコースを学ぶことの潜在的利益を積極的に宣伝しなければならなくなっている。

- デジタルキュレーションに特化した大学院コースを教えようと試みた米国のライブラリスクールは、学生のための適当な実務研修先を探そうとして問題に遭遇した。純粋なデジタルキュレーション要素を持つ職場を機関内で見つけることは、たとえ存在したとしても、非常に難しい可能性がある。デジタルキュレーションを行う図書館や機関の数が増えない限り、実務研修先の不足が教育・研修プロセスの隘路になり続けるだろう。

- デジタルキュレーションの内容を持つコースを教えた経験を持つライブラリスクールは、事態を好転させるために、学生は少なくとも将来働こうと考えている学問分野において学位取得レベルの基礎を持つ必要があると決定した。デジタルキュレーションの役割を果たすべく採用されるライブラリスクールの卒業生は、データだけでなく、一緒に働くことになる研究者の仕事内容も理解しなければならない。明らかに、この条件は学生の勧誘や獲得を難しくさせる。図書館員はこれまでも主題専門家としての役割を切り開いてきたので、同じ事をデジタルデータ分野においてもできるに違いないと述べた者もいた。しかし、ここ数年の傾向として、主題図書館員の役割は、教官と連携して教官のニーズを満たすことに焦点を絞ったリエゾン図書館員に取って代わられつつあるように思われると言う者もいた。研究者のニーズにはコレクション構築の問題やスキルアップ教育などが含まれる場合もあるが、これは、特別に深いレベルの主題知識を意味したり、要求したりするものではないからである。

6. 考察

現在はデータ科学の黎明期である。英国には、データキュレーションセンターのデータ管理者フォーラムなどのイニシアティブの助けを借りて集まり始めたデータ科学の原始コミュニティが存在し、ほとんどコミュニティ形成あるいは専門化という言葉の手前まで来ている。

助成団体や研究コミュニティが研究データのキュレーションや再利用の価値を正しく評価するようになるにつれ、データ作成者と密接に連携して働くデータ科学者は益々重要な役割を持つようになるだろう。理想的には、データ科学者は研究者としての経歴を持つだけでなく、技術的な適性とよく訓練されたアドボカシ能力や対人スキルを持つことになる。これらすべての属性を持ち、さらにデータ科学者としての経験を持つ者は供給不足であり、雇用者はしばしばデータ科学の空きポストを埋めるのに苦労している。

適切な能力を持つデータ科学者の供給に取り組む必要がある。まず、助成団体や研究コミュニティの指導者はデータ科学者が果たす役割の価値を研究コミュニティに大々的に宣伝する必要がある。そして、データ科学それ自体を学術的なキャリアとして認識するという考えを持たなければならない。現時点では、キャリアアップの可能性やインセンティブ、報酬の規模は限られている。

また、デジタルデータをキュレート、保存、保管するデータ図書館員の役割も益々重要になるだろう。これは多くの点で、学術コミュニティに奉仕する図書館の中核的使命の当然の拡張であり、機関リポジトリの研究支援機能から自然に導かれるものである。この分野で働く図書館員は、各専門分野のデータ科学者と密接に連絡を取ることになるだろう。デジタルキュレーションカリキュラムを作成する図書館学教員の作業はかなり進行しているが、現時点では、デジタルキュレーションスキルはライブラリスクールでは広く教育されていない。図書館管理者は、適当なキャリア構造を考案することにより、生まれつつあるデータ図書館学という分野を育てるベストな方法を考える必要があるだろう。

図書館が積極的な立場をとる必要性があるか否かをめぐる議論は、現時点での主要な問題の 1 つを浮き彫りにする。データ主導の研究世界には、研究者自身、図書館、英国研究会議、その他の研究助成団体、IT 関係者、そして最後になったが重要な機関など、数多くのステークホルダーが存在する。現時点では、これらのステークホルダーの中にこの分野の発展を導く指導的な役目を果たしている者はいない。二重支援システムのどちらの側がデータ科学への責任を担うべきかという大きな問題も存在する。

その答えは、どちらの側も責任を免れないということである。当然の如く研究会議の助成を受ける「ビッグサイエンス」は、データ科学者が既に働いているところであるが、専門化するためのシステムや適切なキャリア構造、正当な認識はまだ完全には実現していない。現時点で多くのデータを生産しており、今後さらに多くのデータを生産すると思われる、より小規模な研究プログラムである「スモールサイエンス」は、小規模研究やそのデータを管理する実体は他に存在しないので、研究を行う機関が責任を持たねばならない。すべての分野の研究者にデータを最適に作成・管理するために必要な専門知識が提供されることを保障するために、この機関レベルにおいてもデータ科学者のスキルが必要になるだろう。

それゆえ、研究コミュニティの将来のニーズが解決されるか否かは、研究機関がデータを真剣に考えるか否かに強く依存するだろう。既に検討したように、図書館はデータの保管・保存サービスを機関に提供することにより研究を支援する役割を持っている。しかし、データ科学スキルの提供は機関の研究管理レベルにおいて、計画、実施、管理される必要があるだろう。

そのような措置が非常に少数の機関でしか実現していないことは驚くことではない。データ関連の出来事はこの数年間で急速に進んだので、機関や助成団体が順応する機会はほとんどなかった。体制、プロセス、方針の面でデータ関連の動きについていくことができなかった。

本研究から得た全般的な教訓は、英国のデータ科学は萌芽期ではないとしても、少なくともまだ幼年期であるということである。データセンターなど、スキルという意味でのベストプラクティスの例が存在し、臨床データ管理協会[17]など、役割の専門化という意味でのベストプラクティスの例も存在する。主要な問題は、データ専門家としてのキャリア構造の欠如、研究コミュニティが存在する様々な状況において必要とされる役割とスキル、両者のあいまいな定義、データ科学が進展しうる方法に関する明確なイメージの不在に関するものである。次章で示す勧告がこれらの問題を解決するプロセスを支援することを期待する。

[17] http://www.acdm.org.uk/resources.aspx

7. 勧告

本研究による主要な勧告は以下の通りである。

1. 研究ドメインにおけるデータスキル開発に関する勧告（RD）

勧告 RD1: 英国の主要な研究助成団体は、大学および研究機関と協力して、データ科学者の役割を正しく定義および形式化し、データ科学者の仕事を認識させ、報いることができる方法を策定するべきである。

勧告 RD2: 上と同じ団体は協力して、データ科学を支援し、その研究を促進し、その職務の専門化を促進する条件を作成するべきである。

勧告 RD3: JISC および研究を委託するその他の組織は、次の課題を扱う研究を推進するべきである。

- データ科学者により担われる役割およびデータ科学者が研究に果たす貢献の価値の記述
- データ科学キャリアの例
- データ科学におけるグッドプラクティスを表す一連の実務の開発

勧告 RD4: 関連団体（HEFCE と研究会議）は、データ管理の基礎を扱いそれにより基本的なデータ科学スキルを研究プロセスに組み込む、大学院レベルの短期研修コースを研究者に提供するスキルを持つ指導者のネットワークの構築とそれへの資金提供を検討するべきである。既にいくつかの研究会議は、助成金交付申請にデータ計画の記入を要件とすることで、このための基礎を構築している。

勧告 RD5: 研究会議およびその他の研究助成団体は、助成金交付申請および交付決定プロセスの一部として、プロジェクトチームのうち少なくとも 1 人をプロジェクトのデータ科学者として指名することを要件とするべきか否かを検討するべきである。指名された者にはデータ科学とデータ管理の基礎を提供する短期コースに参加することを必須とするべきである。研究会議は、正当なコースの認定と出席証明が必要とされる範囲を検討するべきである。

2. 研究図書館におけるデータスキル開発に関する勧告（RL）

勧告 RL1: 英国の研究図書館コミュニティは、大学および研究機関と協力して、データ図書館員の役割を正しく定義および形式化し、データ処理スキルを持つ図書館員の適切な供給を保障

するカリキュラムを開発するべきである。

勧告 RL2: JISC は、国際データキュレーション教育活動（IDEA）作業グループの発展を支援することを検討するべきである。このグループは、特に、図書館・情報学を出自とする将来のデータ図書館員のための適当なカリキュラムの作成において重要な諮問的役割を果たす位置にいる。

3. 一般的なデータスキル開発に関する勧告（RG）

勧告 RG1: 既に、データ分野で活動している多くの関係者が存在するので、データスキル研修において相乗効果を有効に活用できる可能性がある。この可能性、特に、UK データアーカイブ、データ科学分野を先導する大学や研究グループ、ライブラリスクール、デジタルキュレーションセンター、IDEA（国際データキュレーション教育連合）の活動を視野に入れた研究を勧告する。この研究は、米国やカナダ、オーストラリアの活動など国際的な調査も行った方が良いだろう。

参考文献

Brown S and Swan A (2007) Researchers' Use of Academic Libraries and their Services (2007). Published by RIN in association with CURL. http://www.rin.ac.uk/researchers-use-libraries

Canadian Digital Information Strategy (2007). Library and Archives Canada. http://www.collectionscanada.gc.ca/cdis/index-e.html

Carlson S (2006) Lost in a sea of science data. *Chronicle of Higher Education*, 23 June. http://chronicle.com/free/v52/i42/42a03501.htm

Greer C (2007) A vision for the digital data universe. Presentation. https://www.nanohub.org/resources/2291/

Henty M, Weaver B, Bradbury S and Porter S (2008) Investigating data management practices in Australian universities. Australian Partnership for Sustainable Repositories (APSR). http://www.apsr.edu.au/investigating_data_management

Lyon, Liz (2007) Dealing with Data: Roles, Rights, Responsibilities and Relationships: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

National Science Foundation, National Science Board (2005) Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. http://www.nsf.gov/pubs/2005/nsb0540/

Martinez L (2007) The e-research needs analysis survey report. CURL/SCONUL Joint Task Force on e-Research. www.rluk.ac.uk/files/E-ResearchNeedsAnalysisRevised.pdf

Steinhart G, Saylor J, Albert Paul, Alpi K, Baxter P, Brown E, Chiang K, Corson-Rikert J, Hirtle P, Jenkins K, Lowe B, McCue J, Ruddy D, Silterra R, Solla L, Stewart-Marshall Z, Westbrooks EL (2008) Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library. A report of the Cornell University Library Data Working Group. May 2008. http://ecommons.library.cornell.edu/handle/1813/10903