Vol.2017-CH-114 No.10
2017/5/13

人文情報学における共創型研究とオープンサイエンスの潮流

北本 朝展[†1†2]

概要：今後の人文情報学の課題として，モニュメント的な大目標を掲げ，多くの研究者が資源や知識を共有する機会を提供し，集団的に解決策を見出していく研究プログラムを作り出すという課題を取り上げてみたい．この課題を取り上げた動機は，個人的な競争型研究を越えた集団的な共創型研究への期待である．近年のオープンサイエンスの潮流においても，研究者集団を対象としたグランドチャレンジ型研究やコンテスト型研究，あるいは市民や企業を巻き込んだ超学際的研究を通して，個人を越えた集団としての研究パフォーマンスをどう向上させるかという視点が欠かせない．そこで本発表では，人文学オープンデータ共同利用センター(CODH)が公開するオープンデータを中心に，人文情報学においてどんな共創型研究が構想できるかを考えてみたい．

キーワード：人文情報学，共創型研究，オープンサイエンス，グランドチャレンジ，くずし字

Co-creation in digital humanities and the trend of open science

ASANOBU KITAMOTO[†1†2]

Abstract: As a challenge of digital humanities in the future, we discuss the issue of creating a research program that proposes a monumental challenge, offers many researchers a chance to share resources and knowledge, and collectively tries to find solutions. Motivation behind this challenge is expectation for collective research based on co-creation beyond individual research based on competition. Recent trend in open science raises interests on improving performance of research collectively than individually through grand challenge research or contest-based research among a group of researchers, or trans-disciplinary research with participation of citizens and companies. Hence this paper discusses the concept of co-creative research in digital humanities, centering on open data released from Center for Open Data in the Humanities (CODH).

Keywords: Digital Humanities, Co-creative Research, Open Science, Grand Challenge Old Japanese Characters

1. はじめに

研究には様々なスタイルが存在するが，個人を主体とするか，集団を主体とするかは，得られる研究成果の性質を左右する重大な問題である．理工系においては集団による研究は一般的な存在であるが，人文系においては個人による研究が主流であることが特徴的である．

この理由として，人文系では研究が伝統的には個人の創造性に基づき行われてきたこと，そしてテキストを読んで解釈するという研究方法に限定すれば複合的な技能を必要としないこと，などを原因として挙げることができよう．しかし後者についてはデジタルヒューマニティーズの広がりに伴い，情報学と人文学との協働が広まって変化が生じる可能性もある．

一方で，研究は集団によって進むという側面は分野を問わずに存在する．その代表的な存在が学会や研究会などであり，コミュニティにおける議論によって知識を共有することは一般的に行われている．さらに共同研究などの形で得意分野を補うことによって，個人ではできない大型のプロジェクトを推進することは，人文学においても一般的に進められてきた．例えば人文科学とコンピュータ研究会（CH研究会）の陣容が整った背景には，1995 年から 4 年間続いた重点領域研究「人文科学とコンピュータ～コンピュータ支援による人文科学研究の推進」が大きく寄与したことが指摘されている[1]．このように共通のテーマで研究者が集うコミュニティを形成することはプロジェクトの重要な意義であり，それによって集団内での知識の共有が進むことになる．

しかしインターネットが普及し，オープンサイエンスの時代に入って，知識の共有にとどまらず知識を共創することを目指した新しい研究スタイルが出現してきた．本論文ではこれをグランドチャレンジ型研究，ワークショップ型研究，コミュニティ型研究と類型化し，その特徴を分析する．さらに人文情報学においてこの種の研究を進めるための一案を提示してみたい．

本論文の構成を以下に述べる．まず第 2 章はオープンサイエンスを手短に紹介する．続いて第 3 章は共創型研究を類型化し，それぞれの類型の詳細を述べる．第 4 章は人文情報学における適用を述べ，第 5 章で本論文をまとめる．

2. オープンサイエンスの潮流

オープンサイエンスとは，様々な学術資料をオープン化することによる波及効果を通して研究方法の変革を迫る動

†1 情報・システム研究機構 データサイエンス共同利用基盤施設 人文学オープンデータ共同利用センター
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems
†2 国立情報学研究所
National Institute of Informatics

きの総称である[2]．論文の場合はオープンアクセス，データの場合はオープンデータと呼ばれるが，オープンサイエンスは単に学術資料をオープン化することのみを目的としたものではない．むしろオープン化を梃子とした波及効果として，壁を越えた参加の促進や公平な研究環境の確保，そして透明性の強化などのテーマを含むものである．さらにそうした研究基盤を持続可能とするための，新しい評価システムや人材育成，超学際的な協働などまでを射程とする．ゆえに，短期的に見れば学術資料のオープン化が前面に出てはいるが，長期的には学術研究スタイルそのものを変革するという概念を含むものである．

オープンサイエンスによるオープン化には，様々な壁を取り払うことが期待されているが，それは個人研究主体という個人の壁をどうするかという点にも関わってくる．オープンな研究スタイルによって，個人研究を越えたパフォーマンスを得ることができるかという問いが生まれる．個人を越えて共に創る研究という意味で，このような研究を共創型研究と呼んでみよう．「共創」と「協力」の違いについて江渡は，協力は 1 つの仕事を共に進め，そこで得られた利益を分かちあうことに主眼が置かれているが，共創は「共通善（common good）」という大きな 1 つの目的に向かって異質な才能が集結することに意義があると述べている[3]．

前者は従来の共同研究のスタイルであるが，後者を成立させるためには多くの人々を集結させる「場」のデザインが重要であり，それを持続的に成長させていくためのルールを備えたプラットフォームや，共通した 1 つの目標に共感するコミュニティを形成することが必要となる．こうした新しい研究スタイルの出現によって，従来には得られなかった研究成果を生み出すことへの期待がある．

こうした研究スタイルはまだ一般的なものとは言えないが，知的生産の中でも特にソフトウェアの世界では，オープンソース活動としてすでに一般化しつつあると言える．大規模なオープンソースソフトウェアプロジェクトとは，一つの大目標に共感した専門家が集まる場と言ってよい．そしてみんなで知恵を出し合って一つのプロダクトに集約していくことで，Linux のような超大規模ソフトウェアが今も継続的に開発されている．また知識やデータを蓄積する活動においても，Wikipedia や OpenStreetMap などの成功例があり，いずれもオンラインの百科事典を作る，フリーの世界地図を作るという大きな目標に共感した多くの人々が協力して一つのプロダクトを作っている．ここでは個人という壁を越えた共創が発生しており，それが個人では不可能な大規模知的生産物を生み出す原動力となっているのである．

このような方法論を研究の世界に持ち込めないかと考えるのは自然な発想であろう．関心や志を同じくする人々が集まり，共に何かを作り上げられないだろうか．そのためには共創型研究とは何かという問いを類型化し，それぞれの特徴を分析する必要がある．

3. 共創型研究の類型

3.1 集団で進める研究

オープンソースなどで成功した共創スタイルを研究に持ち込もうという発想は自然ではあるものの，学術研究の世界ではあまり成功例はないというのが筆者の印象である．それはなぜだろうか．一つの回答として研究は共創よりも競争が重視されるという世界であるからという説明がある．またその競争は個人あるいは小さなチーム単位で行われ，競争に勝利したという成果の計測が研究全体の評価に直結するからということになる．ニールセンもオープンサイエンス革命において，こうした方法論を科学に持ち込もうとする際の最大の障害は論文による評価が主流であるためであり，評価ステムを変えることが最大の課題であると述べている[4]．つまり過度な競争環境にある学術研究の世界に，それとは真逆の共創という考えをそのまま持ち込んでも，それは定着するのが困難であると言わざるを得ない．

もしそうであれば，競争と共創とをうまく両立させるような場をデザインすることはできないだろうか．つまり，競争という仕組みを取り入れつつも，そこから得られる知識をうまく集約できるような場のデザインである．その一つの例として，グランドチャレンジ型研究を取り上げる．

グランドチャレンジ型研究（3.2 節）とは，達成に時間を要するような野心的かつ社会的にインパクトが大きいというモニュメント的な目標を掲げ，その大目標に共感した人々がその下に集結することによって，共通目標に向けた知識の共創を行っていくという方法である．個々の研究者が行う研究は異なるが，目標が共通しているため，そこから得られた知識は共有しやすいものとなる．そして大目標に一歩一歩近づきながら目標を上げていくことによって，全体の知識を段階的にアップデートしていけるという点も魅力的である．

しかしこうした目標は野心的でありすぎると，共感できない人には参加しづらいものとなるため，よほど優れたリーダーシップがないとうまく運営できないという欠点がある．そこで目標としては野心的ではないが，共通の目標を共有するスタイルとしてワークショップ型研究（またはコンテスト型研究）（3.3 節）が広く行われている．ここで設定される目標は，コンテストのために用意された共通のデータセットで優れたパフォーマンスを出すという目標である．ただし評価データと評価尺度が参加者で共通のため，同じ基準で複数グループの研究を相対的に評価できるようになり，知識を集約してそこから共創しやすくなる．

この 2 つの研究スタイルは，競技型の考えを導入することで知識の集約を容易にするという場のデザインを採用しているが，これは数値的に優劣がつけやすい研究テーマに

適した方法であり，いつでも適した方法というわけではない. 特に人文学の問題では, 複数の解釈のどれもが正しい，あるいはそもそも何をどう解釈するかという基準そのものが研究の対象という場合が多く，それをお仕着せの基準で比較するということ自体に拒否反応が生まれることもあるだろう．そうした場合には競技という方法を経由せずに共創を生み出す場のデザインを行う必要がある．それがコミュニティ型研究（3.4 節）であり，そこでは合意した目標に向かって人々が共に創造していくプロセスを作り出す場をデザインする．

3.2 グランドチャレンジ型研究

グランドチャレンジ型研究とは，実現できるかどうかもわからない大目標を掲げ，その実現に向けて多数の人々が知恵を出し合うことで，目標に一歩ずつ近づいていくという方法である．有名な例に，米国の DARPA が開催する DARPA Challenge がある．例えば自律走行車の競技は，自律走行がまだ夢物語であった 2004 年に, 砂漠を使った無人走行車の技術を競うコンテストとして開催された．このような大目標に関心を持つ研究グループが同じ場で競い合うことで，自律走行に向けた研究課題の共有や人的交流が進んだ. それから 10 年が経って自律走行はブームを迎えたが, その裏ではこのグランドチャレンジ型研究が大きな役割を果たしていると言えるだろう．

もう一つの代表的な例にロボカップ[a]がある．これは 2050 年までにロボットチームが人間チームにサッカーのゲームで勝利することを目標としたプロジェクトであるが, 最終的なゴールは遠いことから段階的にレベルアップしていけるように競技を設定することで，その時代の技術レベルで競技に参加しながらレベルアップしていけるように設計されている．ここで重要なのが，期限を切ることの重要性である. ロボカップでは当初は目標達成時期を 2100 年ごろに想定していたそうだが，アドバイザーに前倒しするように言われて 2050 年としたらしい．このように目標が 50 年も変わりうるということは，むしろ期限を宣言することに意味があると考えた方がよい．

日本で注目を集めた「ロボットは東大に入れるか」[b]プロジェクトも, 2016 年度までに大学入試センター試験で高得点をマークする, また 2021 年度に東京大学入試を突破するという形で期限を決めている．とはいえこれらの目標は，あくまで目標を具体的に描くことでタスクを明確にするための象徴的存在に過ぎず，東大入試を突破するという「実績」を得ること自体が目的というわけではない点に注意したい．つまりグランドチャレンジにおける目標とは，人々の関心を一点に集中させることで参加者の間で相互作用を生み出し，知識を集約するためのツールであるとも言えるだろう．

a) http://www.robocup.org/
b) http://www.21robot.org/

期限の明示がインパクトをもたらすという意味では「シンギュラリティ」にも類似した面がある．人工知能の脅威というメッセージが人々に真剣に受け取られたのは，その時期を 2045 年頃と明示したことにあると筆者は考える(その後，シンギュラリティの到来時期は，2029 年に前倒しされたようである). シンギュラリティ論の裏にも汎用人工知能という大目標があり，そのテーマには多くの研究者が挑んでいる．これも広義のグランドチャレンジ型研究と言えるかもしれない．

グランドチャレンジのように明確な目標を掲げることは, 市民からの協力を得る市民科学型研究においても重要である．市民がそのプロジェクトに協力したいと思うには，そのプロジェクトが世のため人のためになる大義を体現している必要があるだろう．逆に成果が何のために使われるのかが不明確，または個人的な興味に基づくものであれば，協力したいとは思わないかもしれない．つまり大きな目標を掲げることは，研究という壁を越えて広く世界から支援を受けるためにも不可欠の要素と言える．

こうしたグランドチャレンジ型研究は，目標設定が魅力的であるため，人々を惹きつけやすいという点にメリットがある．またグランドチャレンジの目標は遠い将来の話なので，バックキャスティング的に取り組むべき課題を明らかにしやすいという点もメリットと言える．一方で，人々が共有できる魅力的なチャレンジを設定することは簡単ではなく，その夢のような目標を統率し続けるリーダーシップも必要である．また，グランドチャレンジ自体は魅力的であっても，それに向けて現時点で取り組めるテーマがつまらないものになってしまうと，全体の魅力も色あせてしまうだろう．長期的なテーマの魅力をその時々の現実的なチャレンジにいかに落としこめるか，このタイプの研究ではそこが問われることになる．

3.3 ワークショップ型研究

ワークショップ型研究とは，共通の評価データと評価基準を用意することで，研究の相互比較を通して知識の集約と共創を狙うタイプの研究スタイルである．ここで掲げる目標は，グランドチャレンジのような気宇壮大なものではなく，共通の評価基準の中で優れたパフォーマンスを得ることである．そしてパフォーマンスを相対比較するという場のデザインが競技での勝利に向けた動機づけを担い，そこから得られた知識はコンテストでの優劣という基準を参考にして淘汰されることになる．

ただしワークショップ型研究におけるパフォーマンスの優劣は，本来は絶対視すべきものではない．ワークショップ型研究で用意される評価データは実世界問題で扱うべきデータを代表している保証がなく，コンテストで優秀な手法が実世界でも優秀なことを保証することはできない．したがってここで用いられる評価基準は，相対比較を通して知識を集約するためのツールに過ぎず，それ自体を目標と

すべきものではない．ゆえに，コンテストでの成績を手法の優秀さのエビデンスとして活用することは望ましくないとされる場合があるが，ランキングが発表されれば上位手法に注目が集まるのは人情であるのも確かである．そして注目を集めたアイデアが多数に吟味されることで速やかに有用な知識が拡散するというダイナミズムは，ワークショップ型研究の重要な特徴である．

その特徴が最も象徴的に発揮された例がImageNet Large Scale Visual Recognition Challenge (ILSVRC)[c]という一般物体認識に関するワークショップの事例である．このワークショップは，共通データを用いて一般物体認識アルゴリズムの性能を競うことを目的とする．一般物体認識は困難な問題であり，誤認識率は緩やかにしか改善しないと参加者が思い込んでいたところ，2012 年にいきなり誤認識率を10%程度も下げるアルゴリズム[5]が発表され，参加者の度肝を抜いた．これが深層学習（ディープラーニング）と呼ばれるアルゴリズムだったのである．このパフォーマンスは従来型アルゴリズムの細かい最適化では到底達成できないことから，参加者は一斉にディープラーニングの研究に転換し，それが今日に続くディープラーニングブーム，さらには人工知能ブームのきっかけとなった．ワークショップのように共通データセットを用いて実験していれば，その性能がいかにすごいかは一目瞭然である．したがって共通データによる研究は，成果を理解しやすく，新しく生まれた知を迅速に広めることに効果があると考えられる．

また日本でも情報検索を対象としたワークショップであるNTCIR[d]が 1999 年以来続いている．先述のILSVRCもそうであるが，こうしたワークショップでは研究分野として一般物体認識や情報検索を扱うという点は不変であるが，どの評価データと評価尺度を用いるかという点は毎回更新されており，ある程度の性能が達成できた段階で難しい問題や未解決問題にアップグレードするという形で研究の新陳代謝が進んでいる．

しかしこの方法にも，共通の評価データと評価尺度を用意するのにコストがかかること，いったん用意されると研究が過剰適応する傾向が見られること，さらに何も考えずに指標だけに最適化する研究が表れるなどの問題がある．結局のところ，単にベンチマークデータを使うだけに終始し，そこから得られた知見がワークショップという場で共有されなければ，それは集合的な研究スタイルから個人的な研究スタイルへの回帰であり，場を設定していることの意義が失われてしまうという問題がある．その他の問題については関根による論考[6]が多くの貴重な提言を行っている．

3.4 コミュニティ型研究

そもそも学会とはコミュニティであり，学会とはそこに人が集うことによってインタラクションが生まれ，それによって知識が共有され生まれることを期待する場である．その点では，そもそも学会とは集団による研究の中心となるべき存在と言えるかもしれないが，現状の学会の機能はむしろ知識の権威付けや組織化に移っており，場のデザインとしては共創に向いているとは言えない．講演形式の発表は深いインタラクションには向いていないため，表面的な知識の共有はできるとしても，それを集約して新しい知識を共創するに至ることは困難である．

一方，集団で進める研究として従来から存在するのが共同研究という仕組みである．これはリーダーが大まかに研究テーマを設定し，その要素となる研究を参加者が分担して進め，最終的に得られた部分的な成果を集約して全体の成果にまとめなおすという方法に基づく．分担という方法を用いることで，幅広い研究テーマが扱えるようにはなるが，グループごとに閉じた体制で研究が進むことが多く，最終的な成果は並置というレベルを越えることが難しい．共同研究グループに所属することによって多少のインタラクションは生まれるものの，マネジメントの困難さもあって，多くの場合は個別研究の寄せ集めという形態にならざるを得ない．

このように，学会や共同研究という従来型の研究スタイルの限界を踏まえて，共創というレベルで新しい研究スタイルが可能な方法に向けた模索が始まっている．例えばハッカソンやアイデアソンなどの○○ソンと呼ばれるイベントはその一例である．「ソン」とはマラソンのソンであり，ハック＋ソン，アイデア＋ソンなどの造語が様々に生まれているが，いずれも共創を目指した場のデザインに工夫がある．具体的には，参加者同士のインタラクションから新しいアイデアが生まれるのに十分な時間を費やす，未成熟と思えるアイデアであっても批判しない，すべての人が何らかの形で参加できるように配慮するなど，デザインとルールを適切にコントロールすることで参加者の意識を高めて創造力を引き出すことを目指している．

しかし，こうしてボトムアップに共創された知識に持続性を持たせることが別の課題となる．コミュニティ型研究では場のデザインを重視するあまり，生まれる知識の質を高めるよりも場の雰囲気を盛り上げることの方に気が向いてしまい，イベントとしては成功だが結果としては失敗ということにもなりかねない．また最初からインタラクションがオープンになってしまうため，その場に出たアイデアや知識が盗まれやすいという問題もある．さらにこうした場が「安くアイデアが得られる場」として運用されてしまうとアイデアを搾取する場となってしまい，持続性のないソリューションとなってしまう危険性がある．

そこで重要となるのが参加者にも利益が生まれる場のデザインであり，そこで重要となるのが市民科学（シチズン・サイエンス）の考え方である．この場では市民は単なる労

c) http://www.image-net.org/challenges/LSVRC/
d) http://research.nii.ac.jp/ntcir/index-ja.html

働力（クラウドソーシング）ではなく，その場に参加することを通して市民自身が問題を考えて知識やスキルを得るという主体的な存在である．例えば「みんなで翻刻」[e]においては，その場に参加する市民は翻刻データを増やすという役割を担うだけでなく，翻刻作業に参加することによってくずし字を読む能力が向上するという利益も得る．さらに翻刻作業の難しい点を上級者と議論することにより，参加する市民は教育的な指導を受けつつより深い理解に到達することもできる．このように研究者にも市民にも利益がある場をデザインするということが不可欠である．

このようにコミュニティ型研究にも様々な試みがあるが，まだ全体に未成熟なところが多く，どのように場をデザインすれば成功するのか，まだまだ未知の部分が大きい．いくつかの成功例を注意深く分析することによって，成功に寄与すると考えられる要因を特定し，それを一般化していく研究が望まれる．

3.5 個人と集団の関係

以上で 3 種類の共創型研究を比較してきた．全体をまとめると，グランドチャレンジ型研究は目標が中心，コミュニティ型研究は場が中心，そしてワークショップ型研究はその中間として競技を中心にすると言えるだろう．目標を中心にする場合は，磁力の強い目標に人々がひきつけられることによって，自然にあるいは必然的に場が形成されるというトップダウン的なスタイルが中心となる一方，場を中心にする場合は，場をデザインすることによってインタラクションが発生して新しい知識が創造されるという，ボトムアップ的なスタイルが中心となる．ワークショップ型はその中間として，目標と場を比較的小規模にとどめることによってデザインの複雑さを軽減するとともに，小目標を競技で達成する場を不断に更新することで，長期的に持続可能なデザインを確立していると言える．

このように各種のアプローチを類型化することで，自分が実現したい研究がどのスタイルに合っているかを判断し，それに応じた設計をすることができるだろう．

4. 人文情報学における展開

4.1 アイデアソン

人文情報学の分野で，このような共創型研究を展開できるだろうか. まずは筆者自身が体験した小さな実例として，アイデアソンから誕生したアイデアを現実化させた例を紹介したい．「江戸料理レシピデータセット」[f]は，筆者が日本古典籍に関するアイデアソンに参加した際に，江戸の料理本を見て「これを現代のレシピに翻訳してクックパッドに掲載したら面白いのではないか」とアイデアを思いついたときから始まった[7]. 筆者はこのアイデアソンに参加するまで，このような料理本の存在を全く認知していなかった．しかしアイデアソンで議論を重ねてアイデアがまとまったので，それをクックパッドに持ち込んで会社の担当者からも賛同を得て，データ作成などの行動を開始することで, およそ 1 年後にアイデアを現実化することができた．

これはアイデアソンというコミュニティ型研究においてそれまでは考えもしなかったアイデアが生まれ，それが新たな可能性を開いたという例である．ただしこれはたまたま成功した例であり，いつでも繰り返せるものではない．アイデアソンもあくまでツールの一つであり，他のツールも含めて幅広く可能性を追究する必要がある．

4.2 グランドチャレンジとしての電子テキスト化

そこで考えるのが，電子テキスト化というグランドチャレンジである．人文学においてテキストは特別な重みをもつものであるが，それを電子的に扱うためには本を電子テキストに変換しなければならない．もし電子テキストになれば，人類が生み出したこれまでの文化を網羅的に共有できるだけでなく，それをコンピュータ（人工知能）に分析させることも可能になる．これは人文学の夢でもあるし，社会的なインパクトも大きい．だから世界中の本をまずデジタル化しよう，そう考えて始まったのが Google Books であり, その他の大規模書籍デジタル化プロジェクトである．しかし，本を画像にデジタル化したあと，さらに電子テキストに変換しないと簡単に読むことはできない．その最後のステップをどう解決するかが大きな問題として残っている．その役割を担う主体には 2 つの可能性がある．人間か機械である．

もしこのステップを人間が担うのであれば，業者によるアウトソーシング，市民によるクラウドソーシング，ボランティアによる分担作業などを通してテキスト化を進めることになる．クラウドソーシングについては，University College London におけるBentham Project[g]などは大規模なテキスト化を進めているし，日本では「みんなで翻刻」プロジェクトがくずし字の翻刻を進めている．ボランティアによる分担作業については，世界的にはProject Gutenbergなどが著名であるし，日本では青空文庫が代表的である．人間が結果を確認することで，信頼性の高い電子テキストを生み出せるという点が，この方式の最大のメリットである．

一方，このステップを機会が担うのであれば，光学的文字認識（OCR）ソフトウェアの開発が不可欠である．例えばGoogle BooksはOCRでテキスト化されているし，その結果はHathiTrust Digital Library[h]などでも利用できる．しかしGoogle BooksのOCRといえども言語によっては精度にばらつきがあり，まだまだ万能とは言えない段階にある．したがって文字列やN-gramを用いた検索には有用であるとしても，そのまま研究に利用できるレベルのテキストは得

e) http://www.honkoku.org/
f) http://codh.rois.ac.jp/edo-cooking/
g) https://www.ucl.ac.uk/Bentham-Project
h) https://www.hathitrust.org/

られないと考えた方がよい．

4.3 グランドチャレンジとしての OCR 開発

OCR の研究には長い歴史があり，ある意味ではすでに研究し尽くされた分野である．その成果として，OCR にはオープンソースから商用製品まで様々な選択肢が用意されており，OCR に適した本であれば比較的高い精度を簡単に得ることができる．ただしそれは現状の OCR に適した本は読めるというだけであり，それでは読めないという本がまだ大量に眠っているのが現実である．

例えば現在最もよく使われているオープンソースソフトウェアであるtesseract[i]に注目してみよう．これはもともとヒューレットパッカード研究所にて 1985 年から 1994 年にかけて開発されたソースコードを源流とするソフトウェアで，2005 年にオープンソース化され，2006 年からはGoogleが開発に関与，2017 年にも新しいバージョンをリリースするなど今も開発は活発に続いている．その理由はGoogleのサービスのいくつかでOCRが利用されているためで，Googleは今でもOCRの精度を向上させることに意欲を持っている．

さらに近年のディープラーニングの発達が，OCR 開発に再び火をつけている面もある．Tesseract は現在のバージョン 3 までは従来の画像処理手法を用いていたが，バージョン 4 からはディープラーニング（LSTM）ベースのエンジンに置き換えられる予定で，2017 年 4 月現在はアルファ版がリリースされている．開発者の報告によると，ディープラーニングの導入によって誤認識率が数割程度は減少するとの実験結果が得られているとのことで，ここでもディープラーニングによる画期的な精度向上が見込めると考えられる．このように，研究としては終わったと思われる OCR の分野であっても，まだやるべきことはたくさん残っているのである．

歴史的な本のOCRにtesseractを使うというプロジェクトもいくつか立ち上がっている．例えばテキサスA&M大学が進めたEartly Modern OCRプロジェクト[j]は，tesseractをOCRエンジンとして用い，独自に開発した周辺ツール群に英国サルフォード大学の開発ツールを加えたワークフローを構築することで，初期英語書籍を中心としたテキストのためのOCRシステムを構築した．またChinese Text Project [k]は，tesseractを独自に改良することで古い中国語のOCRに成功し，このOCR結果を修正するインタフェースも提供することで，研究にも利用できる電子テキストデータベースを独自に運用している．

このように人文学の立場から見れば歴史的な本の OCR が課題として残っており，そこはビジネスになりづらいことから商用製品では開発しづらい領域である．しかし人類の文化を網羅的に分析するという夢を実現するには，歴史的な本を電子テキスト化するというのは大きなチャレンジである．したがって OCR 開発を中心とする機械の発達と，市民参加を中心とするプラットフォームの発達とを組み合わせて歴史的な本を電子テキスト化するということは，人文情報学における一つのグランドチャレンジと捉えることができる．

4.4 くずし字チャレンジ！

このように残された課題の一つに，日本古典籍のくずし字がある．おりしも国文学研究資料館を中心とする「歴史的典籍NW事業」[l]が日本古典籍 30 万冊のデジタル化という大きな事業を進めており，これから数年の間にこの事業からは大量の画像が生み出される見込みである．またこのうち一部は，国文学研究資料館との共同研究として，情報・システム研究機構 データサイエンス共同利用基盤施設 人文学オープンデータセンターのウェブサイトから，「日本古典籍データセット」[m]としてダウンロード可能なファイルとして公開している．また，画像に加えて一部の古典籍に関しては，「日本古典籍字形データセット」[n]として翻刻の過程で生み出される文字情報を整備しており，Unicodeと文字のバウンディングボックスがセットになったCSVファイルを公開している．ゆえにこれらの本に関しては，画像に何が書かれているかを人間が読むだけでなく，コンピュータに学習させることも可能である．

しかし一般的には，画像があるだけでは本の中身にアクセスすることができないため，1 ページごとに読むしか本の内容を把握する手段がない．そこで本の中身を検索するための「ディープアクセス技術」の開発が必要であり，そのうち特に文字に対するアクセス技術としての「くずし字 OCR」や「歴史的文書 OCR」の開発が重要課題となる．

くずし字の機械による解読に向けた OCR 研究にはすでにいくつかの例があるが，大量の画像を自動的にスキャンして電子テキストを作ってくれるソフトウェアはまだ存在しない．したがって，このようなことが可能な OCR ソフトウェアを開発するというのが第一のグランドチャレンジと言えるだろう．

その完成の期限を決めるとするならいつになるだろうか．そこで参考にするのがヒトゲノム解析の歴史である．ヒトゲノムの解析は初期には 100 年かかる（＝実質的には不可能に近い）と言われていたが，機械による自動的な配列決定技術が提案され，その大目標に向けた技術開発が驚異的なスピードで進んだことから，アイデアの提案からわずか 15 年程度でヒトゲノムの全解読が完成した．さらにその後も技術の進歩が進み，当初はヒトゲノムの解読に 3000 億円もかかったのが，今や数十万円から数万円に向けて価格が

i) https://github.com/tesseract-ocr
j) http://emop.tamu.edu/
k) http://ctext.org/
l) https://www.nijl.ac.jp/pages/cijproject/
m) http://codh.rois.ac.jp/pmjt/
n) http://codh.rois.ac.jp/char-shape/

下落する過程にある．このように，技術進歩が積み重なることにより，当初には想像していなかった形でグランドチャレンジが達成されることがある．

同様に歴史的な本に書かれた内容を網羅的に解読するという「スクリプトーム解析（scriptome analysis)」が完成するのはいつ頃になるだろうか．上記のヒトゲノム解析の事例を参考に，今から 15 年後の 2032 年頃としてはどうだろうか. そうすると，その 15 年間になすべき課題をバックキャスティング的に洗い出して，研究を段階的に進めることが可能になる．ちなみにシンギュラリティの予測もほぼ同じ時期であり，2030 年頃というのが一つの目安として妥当なのではないかと考える．

第二のグランドチャレンジは, 人間と機械の協調である．その意味では，今年になって始まった「みんなで翻刻」プロジェクトにおいて予想外のハイペースで翻刻が進んだことは一つの明るい材料である．日本には，翻刻を通してくずし字の解読にチャレンジするとともに，電子テキストの共有に貢献したいと思う人々が案外たくさんいるのかもしれない．そうした人々をさらにうまく組織化して人間による翻刻を進めると共に，さらに機械とお互いの能力を補うことによって能力を高める道筋を提供するというチャレンジである．

そこではアルファ碁によって起こった囲碁界の変化が参考になるかもしれない．つい 1 年前まで，コンピュータが最強の棋士に勝つとのはまだまだ先のことと思えたのだが, 急速に強くなったアルファ碁は最強の人間にあっさりと勝ってしまった．しかし勝敗そのものよりも衝撃的だったのが，アルファ碁がとった戦略である．アルファ碁は人間には想像できない戦略をアルファ碁同士の自己対戦によって編み出し, それを実戦で使って勝ったのである. 対戦前は，ソフトウェアの強さは詰碁的な局面の後半戦にあると想像されていたのだが，実際には前半戦における布石という人間的と思われていた局面における強さが圧倒的であった．これに衝撃を受けた棋士は，むしろ「アルファ碁先生」から新しい戦略を学ぼうと頑張っており，コンピュータが生み出した戦略から新しい囲碁の世界が広がろうとしている．同じようにくずし字を対象とした人工知能が十分に発達すれば, コンピュータから学んだくずし字の読み方によって，さらに多くの人々がくずし字を読めるようになる可能性もある．このように，人間と機械が協調しながらくずし字文化を広めるシステムを作ること，これがもう一つのグランドチャレンジになるだろう．

以上のようなグランドチャレンジを「くずし字チャレンジ！」と題して，その場に人々の知恵を集めて共創することはできないだろうか，というのが本論文の提案である．

もちろん，グランドチャレンジのテーマはこれに限ったものではなく，他にも考えられるかもしれない．また，本論文で挙げた 3 種類の共創型研究のいずれも人文情報学には適用可能である．つまり，そうした共創型研究を人文情報学に導入することによって，研究分野としてまとまった成果を生み出していき，他分野あるいは社会にも見せていくことが必要ではないか，というのがより広い視点からの提案である．

5. おわりに

個人ではなく集団による研究でパフォーマンスを向上させるという研究スタイルに関する様々な試みをまとめた．このように大きな目標を決めて知恵を集約するという研究スタイルは，何かを新しく作り出すというイノベーション型研究と非常に相性が良い．その意味で，人文情報学のうち情報学の部分においては，こうした研究手法を考えることに意味がある．

一方，人文学における研究テーマは，必ずしもこうしたイノベーションに関連するものではない．むしろ個人的な興味に基づき，限られた範囲のテーマを深く掘り下げることによって成り立つところが多く，共通目標を設定すること自体が馴染まない場合も多い．そのような場合に共創という言葉はピンとこないかもしれない．

そこで，人文学における大目標を設定し，それを実現するための情報学の技術開発のマイルストーンを設定するというのが，本論文の提案である．具体的には，本の中身へのディープアクセス技術のうち，特に文字に焦点を合わせた OCR の開発と，人間と機械の協調によるくずし字電子テキスト化の進展である．その目標が多くの人を引き付けるほど魅力的なテーマであるとは断言できないが，少なくとも人文学から見れば重要な課題であることは確かである．そこに知恵を集約してモニュメント的な成果物を作ることができるか，本当に難しい目標としてグランドチャレンジにふさわしいテーマであると言えるだろう．

参考文献

[1] 相田 満.人文科学とコンピュータ研究会（CH）． 情報処理, 2007, vol. 48, no. 6, pp. 664-665.
[2] 北本 朝展. オープンサイエンスの動向と情報学分野へのインパクト. 電子情報通信学会技術報告, 2016, vol. 116, no. 259, pp. 1-6.
[3] 江渡 浩一郎＋くとの. ニコニコ学会βのつくりかた. フィルムアート, 2016.
[4] マイケル・ニールセン. オープンサイエンス革命. 紀伊國屋書店, 2013.
[5] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
[6] 関根 聡. 自然言語処理におけるベンチマークと研究 : エラー分析ワークショップを通じて. 人工知能学会誌, 2016, vo,. 31, no. 2, pp. 269-274
[7] 北本 朝展, 山本 和明. 人文学データのオープン化を開拓する超学際的データプラットフォームの構築, 人文科学とコンピュータシンポジウム じんもんこん 2016, pp. 117-124,