「2014 公開ワークショップ デジタル・ヒューマニティーの最前線と経済学史研究」（2014 年 8 月 25 日、於東京大学大学院経済学研究科・小島ホール 1 階第 1 セミナー室）参加記

森 脇 優 紀

当ワークショップを主催する「デジタル資源を活用した A・スミス経済思想の多元的学際的構造分析の新たな試み」研究班（日本学術振興会科学研究費補助金・挑戦的萌芽研究, 課題番号: 26590031, 代表者:小野塚知二・東京大学・教授）では、東京大学経済学図書館所蔵の「アダム・スミス文庫」について、近年の Digital Humanities（以下 DH）の成果に基づき、Web 上で学術利用し得るようなシステムのパイロットケースを提示し、経済学史研究の情報基盤整備を進めることを一つの目的として活動している。そこで、最近の DH 研究の潮流を知り、その具体的な研究成果を当該研究プロジェクトに活かすべく、本ワークショップが企画された。なお、筆者の専門は歴史研究であり、DH 研究の恩恵も多大に受けているため、その動向には非常に関心を持っている。以下、筆者の関心に沿って内容を報告する。

まず、永崎研宣・東京大学特任准教授（人文情報学研究所主席研究員）が「人文学におけるクラウドソーシングのインパクト：国内外のデジタル・ヒューマニティーズの事例を通じて」と題し、DH 研究の近年の動向とその中で展開されているクラウドソーシングプロジェクトの具体的事例を紹介した。

DH は、1940 年代にデジタル技術を人文学研究に活かそうとする潮流の中で生まれた研究手法で、日本では 1950 年代から用いられるようになった。デジタル技術が著しく進歩した現在では、DH 研究の各団体が Alliance of Digital Humanities Organizations を結成するなど、大規模な組織を形成している。

最近の DH 研究者の間では、“Methodological Commons”が提唱され、各分野が互いの方法論を比較・検討しつつ共有することで、各研究分野においてデジタル技術をどのように用いるべきかを検討し、新しい研究成果につなげていく活動が行われている。こうした流れの中で、University College London（以下 UCL）の Bentham Project（以下 BP）は Transcribe Bentham（以下 TB http://blogs.ucl.ac.uk/transcribe-bentham/ アクセス日:2014 年 9 月 17 日、以下同）を展開することとなった。

人文学におけるクラウドソーシングにおいては、Web 上でテキストデータをデジタル翻刻し、タグ付けをすることが有効であり、不特定多数が一様の方法で資料を扱い、作業したものを共有してチェックし合える点が特徴である。デジタル翻刻にあたっては、人文学の多様な資料・情報を共有するために定められたルールである Text Encoding Initiative（以下 TEI）ガイドラインを採用している。ただし、日本語の特性に対応できていないことから、日本では TEI の知名度は

低く、現在は複数の機関で日本語対応に向けた議論が進められている。

クラウドソーシングによるデジタル翻刻の事例として、TB のほか、米国公文書記録管理局の ”National Archives Transcription Pilot Project”（http://blogs.archives.gov/online-public-access/?p=7171）、ニューヨーク公共図書館の ”What’s on the menu?”（http://menus.nypl.org/）、アイルランドの “The Letters of 1916 Project”（http://dh.tcd.ie/letters1916/）がある。国内には、「翻デジ 2014」（http://lab.kn.ndl.go.jp/dhii/omk2/）があり、日本デジタル・ヒューマニティーズ学会が中心となって、国立国会図書館と国立情報学研究所の協力を得ながら、国立国会図書館の近代デジタルライブラリーで公開されている文献についてテキストデータ化している。

また、DH には Crowd4u プロジェクトに代表される分業化や、Transmedia のような画像化処理の技術も利用されることがある。

こうしたクラウドソーシングプロジェクトは、基本的に人手による作業ではあるが、それがもっとも正確で有効といえる。ただし、適切なフレームワークの確立と、現実的なワークフロー、例えば、デジタル翻刻したものをチェックする人員とチェックするプロセスをいかに確立するかが、不可欠であると永崎氏は指摘する。

永崎報告を承け、DH におけるクラウドソーシングの代表例として、UCL のフィリップ・スコフィールド教授（Professor Philip Schofield）、ティム・コーザー博士（Dr Tim Causer）、クリス・グリント博士（Dr Kris Grint）が、”tranScriptorium and Transcribe Bentham : How to succeed with scholarly crowdsourcing”と題し、UCL に設置された BP の概要と、BP が運営するプロジェクトとして、ベンサムの手稿資料をクラウドソーシングでテキスト化する TB の具体的な作業内容と課題について分担して報告した。

BP は 1959 年に設立され、ベンサムの著作や主として UCL の図書館が所蔵する彼の手稿を集成した”The Collected Works of Jeremy Bentham”の刊行を目指した。その際 BP は、手稿資料を正確に翻字することから始めたという。1985 年より作業にコンピュータを使用するようになり、2010 年に TB が開始された際にはすでに 20,000 葉がデジタル翻刻されていたという。

TB は 2010 年 4 月に始まり、同年 9 月に一般公開された。TB では、ヴォランティアはデジタル翻刻の経験の有無を問われず、ベンサム関係の手稿資料を翻字し、エンコーディングする複雑な業務を担う。データは、テキストの正確さとタグ付けの一貫性のチェックを受け、承認されれば編集できないようにロックがかけられる。データに重大な欠陥があったり、作業が部分的であったりする場合には、そのデータは編集可能な状態に保たれる。

作業実績は、2014 年 8 月までに、部分的なものも含めて 10,195 点の文書がデジタル翻刻されている。不特定多数のヴォランティアによる翻字ではあるが、そのうちの 92%が要求された品質を満たしていたようである。こうした実績の背景には、TB のプラットフォームである Transcription Desk の開発と改良がある。さらに、大英図書館が所蔵するベンサム家の書簡等の手稿資料が利用可能になったことから、資料への関心が高まり、ヴォランティアの数も確保することができているようだ。

一方で、不特定多数の参加が可能であることから生じる問題もある。TB の登録アカウント数は 10,000 あるが、実際に機能しているのは、そのうちの半分か三分の一である。また作業には 440 名が参加したが、その三分の二は 1 点をデジタル翻刻しただけで終わっている。これは、翻字

とエンコードの作業がいかに難しいものであるかを示している。また、歴史的手稿資料の翻字に精通したスーパー・トランスクライバーと呼ばれる25名のヴォランティアが全体の96%をデジタル翻刻しており、彼らに依存せざるを得ないのが現状だという。こういった状況から、TBが真のクラウドソーシングと呼べるまでには至っていないと、TBのメンバーである報告者も自覚しているようである。

TBは現在、EUから資金提供を受けて歴史的手稿資料のデジタル翻刻を行うプロジェクトtranScriptoriumに参加し、ギリシャやスペインの情報科学者、オーストリアのアーキヴィスト、ドイツの辞書学者などと協力して、ベンサムの手稿資料を読み取り、TBに反映させるシステムを開発している。

tranScriptoriumでは、OCRではなく、Handwritten Text Recognition（以下HTR）の技術を用いている。HTRは、単語個別の文字解析やノイズの消去、書入れ線の検出と分別のほか、文字単位で分析する文字検索や単語の予測が可能である。また、隠れマーコフモデルを応用し、これまでに蓄積されたデータに基づいて、原文と一致する文字列のイメージを抽出し、最適のものを提示することができる。こうした技術によって、ユーザーは自動翻字されたものを修正・編集することとなる。ただし、こうした手稿資料を自動認識する技術が、TBのデジタル翻刻作業の正確性と効率性の改善策となるのか、ヴォランティア等の経験・スキルを高めるものとなるのか、またヴォランティアの役割が、デジタル翻刻作業から自動翻字されたテキストの修正・編集へと変化することで、参加者の減少を招いてしまうなど、HTRの有効性を認めつつも、TBに導入する上では課題が残されると報告者は指摘している。

筆者は、16・17世紀の欧文マニュスクリプトを研究で扱っており、翻字作業が困難であること、翻字できる人材の不足を日々体感しているため、クラウドソーシングを通してより多くの人が手稿資料に触れてデジタル翻刻に関心を持つこと、そして作業の正確性と効率性を高める技術の開発に期待をしている。ただし、報告者と同様、自動翻字による作業に依拠することにはまだ慎重になるべきと考える。技術開発に加えて、パレオグラフィーのような翻字のための学問や訓練の場をさらに充実させるなど、手動の翻字のレヴェルを上げる方法の模索も同時に必要ではないだろうか。

なお、二つの講演終了後には総合討論が行われ、クラウドソーシングの参加者としてどのような人材を求めるべきなのか、クラウドソーシングに参加する一般人（社会）と大学の研究機関や専門家との関係をいかにして築いていくべきかという内容について議論が交わされた。

【附記】本稿は、メールマガジン「人文情報学月報 / Digital Humanities Monthly」No.038【前編】に掲載されたものに一部修正を加えて再録したものです。

（もりわき　ゆき：東京大学大学院経済学研究科特任助教）