小特集

ICTが切り開く人文学オープンデータの動向

小特集編集にあたって

編集チームリーダー　北本朝展

人文学とは何か，この問いに対して，人間とは何か，あるいは人間の知的活動とはどんなものかを考える学問であるというのが一つの答えとなろう．この問いを考えるために，人間の知的活動の象徴である書物を読み，内容について他者と議論し，自らの考察を深め，その成果を新たな書物として発表するというのが，人文学の一つの研究スタイルであった．一方，人文学の中でも実証的な性格が強い歴史学や考古学，言語学などの分野では，実世界からデータを集め，分析し，それを総合して新たな理論を立てるというデータ駆動型の研究も盛んに進められてきた．アナログ時代にはカードなどのデータ整理ツールが活用されていたし，パソコン時代には小規模なデータベースも一部で活用されるようになった．

こうしたデータ整理の規模を飛躍的に拡大させると同時に，従来の研究スタイルを変革するインパクトを与えているのが研究のディジタル化である．近年は「人文情報学」や「ディジタルヒューマニティーズ」と呼ばれる研究分野が欧州や米国を中心に拡大している．そして情報学者と人文学者の協働に基づき，最新のICTを活用した新しい人文学研究や，人文学ビッグデータを活用した新しい情報学研究など，分野の壁を越えた研究が盛んになってきた．これはオープンサイエンスにつながる研究の方向性でもある．

オープンサイエンスでは，誰でもアクセスできるという意味でのオープン性，誰でも検証できるという意味でのオープン性，誰でも参加できるという意味でのオープン性などの実現を目指す（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，”信学技報，PRMU2016-90, pp. 1-6, Oct. 2016.）．同様の考え方はデータ駆動型人文学研究でも重要である．コーパスやデータセットを誰でもアクセスできるように公開し，公平な比較などに基づく検証可能性を担保し，分野を越えて様々なステークホルダーが研究に参加できるような環境を構築する．こうしたオープンサイエンスの展開をICTが後押しできれば，データ駆動型人文学研究は更に発展する余地がある．

こうした動向を踏まえ，本小特集「ICTが切り開く人文学オープンデータの動向」は言語・音声と歴史・文学に焦点を合わせ，それぞれの分野で進展する人文学オープンデータの現在の状況と今後の展望を述べる．

前半のキーワードはコーパスである．日本語に関するコーパスは，国立国語研究所を中心に構築が進んでおり，山崎氏は現代の書き言葉のコーパス，小磯氏は現代の話し言葉のコーパス，そして田中氏は現代から過去に至る歴史的な書き言葉のコーパスについて紹介する．研究利用を念頭に，これらのコーパスがいかに注意深くかつ労力をかけて構築されているか，それを知ることによって，コーパスが研究に果たす役割の大きさを実感できるだろう．

後半のキーワードは歴史であり，過去の資料を機械可読とするための文字認識や，人物，地名，時間などのエンティティを軸とした情報の構造化に関する課題を紹介する．日本は世界的に見ても多数の歴史的資料が残っている地域であり，国文学研究資料館のNIJL-NWプロジェクトや各地の大規模ディジタル化プロジェクト等によって，オープンなディジタル画像データの規模も拡大しつつある．こうした人文学ビッグデータに対する人工知能（機械学習）を用いた内容分析への期待も高まりつつある．

過去の日本文化をICTで探るという大いなる挑戦に参加するために，本小特集で紹介したオープンデータに是非気軽にアクセスしてみて頂きたい．

小特集編集チーム　北本　朝展　井ノ上直己　牛久　祥孝　掛谷　英紀
黒川　茂莉　庄司　雄哉　中嶋　秀治　西本　研悟