小シンポジウム「デジタル資源を活用した資料の共有化とこれからの西洋研究への展望」趣旨説明

企画責任者：小野塚知二（東京大学）

はじめに

この小シンポジウムでは、デジタル・ヒューマニティーズ(以下、ＤＨと略記)の技法・手法、それに関わる研究組織や研究環境によって、西洋史学(広く歴史研究の諸分野)にどのような新しい研究の可能性が開かれうるのかを概観したうえで、それらの可能性を現実のものとするために、どのような困難や壁が存在するのかを示し、それらを乗り越えて、いかなる研究組織や研究環境を構築しなければならないのかについて展望を示す。

本シンポジウムで扱うデジタル化された資源は後述のとおり、古い洋書(アダム･スミスの旧蔵書)であり、また、そこに記された手書きの書き込みであり、また、個人の蔵書目録など、基本的に書籍とそれに付随する諸資源であるが、ＤＨで扱うことができる資源は書籍類に限られているわけではない。図像(図版、絵画、写真等々)や音もすでにさまざまなデジタル･アーカイブが構築されており、ＤＨの手法を用いた研究の対象となりうる。

メタデータ

より正確にいうなら、デジタル･アーカイブは、書籍類(より広く、パンフレット、雑誌･新聞なども含む印刷物)と、図像と音の三種類に大別できるというわけではない。デジタル化された資源には、元が何であったかとは別に、必ずメタデータが付与されている。

メタデータとはデジタル化された資源に付された名札か目録情報のようなもので、書籍では図書館情報学でいうところの書誌情報がほぼメタデータにあたる。書誌情報は書籍から自動的には生成しないので[*1]、これまではおもに図書館司書が、蔵書目録を作成する際に、著者名、書名、出版社、出版地、刊行年、判型、頁数、巻数、言語などの情報を書籍から抽出(ないし推測)して、書誌情報は作られてきた。デジタル化以前の書誌情報は、紙(蔵書目録や図書検索カード)に記載されるから、たとえば、西洋近世の書物にしばしば見られる長い書名(『何々についての○○的省察、あるいは◇◇氏の△△に対する××的反論、および●●氏の疑問への包括的な回答；□□博士の序文付き』)は適宜、省略され、またタイプライターで検索カードに打ち込む場合に、ドイツ語や北欧言語のウムラウト記号や、ラテン系言語のアクサン記号、スラブ系言語の子音字への付加記号などは、しばしば省略されてしまうし、中世･近世の綴り方は近代以降の標準的な綴りに書き改められたりもする。デジタル資源の場合、書誌情報は原則としてこのような制約を免れて、より詳細で、より現物に忠実な書誌情報をメタデータとして作成することが可能となる(この点は報告４で詳説する)。

150年前に撮影された写真が原板や印画紙に焼かれた形で残っているとしても、その写真が、いつ、どこで、誰が、どのような機会･状況で、何を撮ったものなのかを示すデータが印画紙の裏面とか、アルバムの余白に書かれていない限り、それは単なる画像にすぎない。100年前の音楽演奏が音源として残っていても、それがいかなる曲目を、誰が、いつ、どこで、誰に向けて演奏したのかといったデータがなければ、それはただの過去の音源データにすぎず、聴いて楽しむことはできるかもしれないが、研究に用いることはできない。印刷物であれ、図像であれ、音であれ、メタデータのないもの、メタデータが不正確であったり、省略されていたりするものは史料的な有用性が低下するのである。

このメタデータがあるおかげで、研究者は書籍類と図像と音という異質の資源の間を横断的に検索･渉猟し、従来よりはるかに多様な種類の資源と大量の材料を用いて研究を展開することができる。た

---

*1　デジタルカメラで撮影された画像にはExif(Exchangeable image file format)という形式で自動的にメタデータが生成するかのように思われているが、それはデジカメがそのように仕込まれているからであって、これも人為の結果である。

とえば、「ラインの護り(Die Wacht am Rhein, The Watch on the Rhine, etc.)」という検索語で、この歌の作詞(1840年)・作曲(1854年)以降、現在にいたるまで、新聞や雑誌、楽譜・歌集、合唱大会や映画・放送、レコードなどに、この歌(1854年以降1945年まで事実上のドイツ国歌)がどのように登場するのかを、とりあえず、ある程度は探すことができる。しかし、現状では、さまざまな資源を横断的に検索するのは必ずしも効率的な仕方ではできない。その理由は、第１に、19世紀末～20世紀前半という時期に関しても印刷物と図像と音のデジタル化は充分には進んでおらず(新聞は殊に英語圏の大新聞ならある程度デジタル化が進んでいるが、その他の言語、また新聞以外の印刷物の本文情報のデジタル化は立ち遅れている)、第２に、デジタル化がされていても不充分なメタデータしか付与されず(殊に図像や音源についてはメタデータが貧弱なためデジタル化されていても、それをある目的から探し出すのに膨大な手間暇を要する)、第３に、多様なデジタル資源のメタデータに統一的なフォーマットが存在していないことにある。

文書の画像データとテクストデータ

ここまで印刷物と図像というように区別して述べてきたが、デジタル資源としては印刷物も手書き文書も諸種の図像もいったんは何らかの画像データとして作成されることが多い。人が文書を読み取るという作業も、実は、文書から直にテクストデータが脳に伝わるのではなく、紙にインクや墨で描かれた記号を目と脳の働きで画像としていったんは認識し、それを各言語に特有の仕方で、文字、文字列、句、文として認識を変換させて、テクストとして認識している。つまりデジタル化以前の文書もまずは画像データとして人に認識されてきたのである。デジタル化によって変化したのは、(1)世界に一点しか存在しない文書を、どこでも居ながらにして読むことができるようになったこと、(2)画像データからテクストデータを抽出し、文書の全文を検索したり、自然言語解析の手法を用いて文書内容を分析(テクスト・マイニング)したりできるようになったこと、(3)画像データを諸種の画像解析の手法で分析できるようになったことである。文書を画像とテクストという両様のデータとして扱うことによって、さらに、(4)手書き文書や書籍への書き込みを翻字(transcribe)した結果をウェブ上で共有・修正でき、また、書誌情報は全く同一だが版による相違を容易に識別できるようになった。五線譜など音楽に関する画像データも音源データと相互対照的・相互補完的に利用可能となるであろう。

このシンポジウムのねらい

本シンポジウムでは、ＤＨによって開拓されるであろうこうした地平をすべて渉猟するわけではなく、近現代の日本が輸入した洋書古典籍の用い方に限定してＤＨの可能性を確認することにしよう。これまでに日本に輸入された膨大な量の洋書には、欧米の著名な研究者や蔵書家のものも含まれており、蔵書票や書き入れなどからは、旧蔵者の思想的背景のほか、当時の社会状況が読み取れることもある。一方、これらを史料として研究者の共有財産にするには、少なくとも三つの課題がある。第１は書誌情報・所蔵情報の公開の問題、第２は書誌情報における歴史的情報の記述方法の問題、第３は歴史資料としての公開方法の問題である。

このうち、第２の課題は、史料の内容読解を重視する歴史研究者の立場から、史料管理に重きを置く図書館や図書館情報学の研究者に対して何らかの問題提起が可能と考えられる。また第３の課題は、洋書のデジタル・アーカイブの必要性、さらには歴史学研究のインフラとしてのデジタル・アーカイブのあり方について、歴史学の側から何らかの行動を起こすべき時期が迫っていることを示している。

この小シンポジムの企画責任者である小野塚知二を中心とした研究チームは、ここ数年、東京大学・経済学研究科所蔵のアダム・スミス文庫を素材として、ＤＨの視点を加味して研究を進めてきた。この小シンポジムでは、これまでの研究の中間報告を行い、前述した２つの課題を解決するための一試論として、デジタル資源を活用した資料の共有化の意味と、これらを活用した西洋史研究への展望について討論してみたい。このシンポジウムは、深貝保則(横浜国立大学)と松尾弘(慶應義塾大学)の両氏を主討論者とし、大澤耕史(東京大学)の司会のもとに行われる。