シンポジウム報告

国文学のデータベース化と著作権

安永尚志†

ご紹介いただきました安永と申します。資料は特に用意しておりませんが、OHPを作ってまいりましたので、それに基づいて説明をしたいと思います。

最初にお断りしておきますが、私は国文学研究資料館という所に所属しておりますが、国文学者ではありません。情報通信工学、とくにコンピュータネットワークを専門にしております。どうして私が国文学と関係があるのかと疑問をもつ方がいらっしゃるかと思いますが、最近は国文学研究資料館でもデータベースをいろいろと作ってきております。そこで、その経歴を紹介すると同時に、本日の話題である製作者という立場から見た諸々の問題、特に著作権に関わるような部分について、問題提起としたいと思っております。

したがいまして、われわれの仕事の紹介と、日頃悩んでおりますいろいろの問題について紹介します。逆に、私から皆様にお聞きしたいと思っておりますので、よろしくお願いします。

本題に入ります前に、シチュエーションを紹介しておきます。国文学研究資料館は文部省設置の国の機関で、大学の共同利用機関の1つです。大学の共同利用機関には、有名な宇宙科学研究所や高エネルギー物理学研究所などがあります。それと同じ組織で、人文科学系の国文学に関する研究所ですが、とくに、専門図書館のような機能を合わせもつ機関です。

人文科学系の共同利用機関には、例えば、国立民族学博物館（大阪）、国立歴史民俗博物館（千葉）、国際日本文化研究センター（京都）などがあります。大学共同利用機関は組織的には大学と同じですが、それぞれの目的を持った言わば目的指向の研究所です。その意味では、業務を抱えているということです。

では、われわれの業務は何かということですが、目的がわれわれにも課せられております。国文学に関する文献資料がありますが、これは一応江戸の末までの日本人が著した著作物、とくに文学作品を対象としています。大体数にすると百万点はあるだろうと言われております。これらは、現在日本中に散在していると同時に、外国にも分散しております。このままでは、研究者が見ることも、たいへん不便な状態にあります。

そこで、文献資料を調査研究し、収集し、整理し、保存するということが重要な役割になって参ります。また、保存しただけではなく、それを広く研究者や一般に提供することが大事な仕事ということになります。したがいまして、その意味から、「国文学の専門図書館」でもあります。

集める資料は、現物が一番良いことは当然ですが、古い本は国宝とか重要文化財とかの文化財であったり、あるいは個人の所蔵になる文化財ですから、なかなか集めにくいという事情があります。そこで、やむなく二次資料ということになるかも知れませんが、写真に撮って集めています。マイクロ資料ということです。マイクロ資料にして永久保存の道を講ずること、これがわれわれの主な仕事ということになります。

次に、国文学とコンピュータのつながりです。コンピュータが、国文学の研究に直接役に立てばよいのですが、まだまだです。そこまで、情報環境は熟成されていないのが実状です。現在のわれわれの仕事は、国文学者が研究する学術情報、すなわち資料と情報をうまく整理して、研究に使ってもらえるようにするということです。これも国文学研究の支援システムと言っております。

本題に入ります。

[OHP] (1) 国文学における学術情報の特質

国文学のためのデータベースをいろいろ作ってきていますが、本日は主題がフルテキストということ

†国文学研究資料館

ですので、まずそれに至る、幾つかの前段を紹介しておきます。

これは「国文学における情報の特質」という表です。国文学で扱う学術情報のいろいろな性質をとりまとめています。特に高次性ということですが、国文学の情報は0次情報から高次情報まで、5つのレベルに分けて考えられます。なぜ0次情報と1次情報を区別しているかと申しますと、0次情報というのは、原文献資料、即ち本そのものに付随する情報です。資料そのものであると思っていただいても結構です。そこから、例えばテキストですと、活字化されたテキストが生成されます。これは、通常本という形で出版されますが、これは原テキストに校訂を加えたテキストということですから、0次ではなくて、1次情報ということになります。したがって、0と1というのはどうしても分けざるを得ないというのがわれわれの基本的なスタンスです。

つぎに、2次情報ですが、これは索引や目録などのように、例えば「文献目録」などと言っている情報ですが0または1次情報をアクセスするための情報です。

3次情報は、1次情報や0次情報の目録のような情報を言います。つまり総合濃縮した情報ということです。

高次情報とは、さらに3次情報よりも広いカテゴリで、例えばある特別の研究動向などを解説したような情報と考えております。

つぎに情報の表現形態ですが、これには複雑な問題が多々あります。簡単に言えば、例えば0次情報ですと、これは文字でも画像でも、あるいは音声として表現されます。そういう意味では多元情報ということです。分かりやすいところでは、1次情報は主として文字で表わします。

[OHP] (2) 国文学データベースの例

これらの情報を、どのようにデータベースとして組織しているかというわれわれの例を示します。0次情報では原文献資料データベースというシステムを作りつつあります。これは原本の各ページをイメージデータとして光ディスクに蓄積し、遠隔地に居る人が利用するとき、例えばファクシミリなどで直接届ける、いわば原文献資料の流通システムです。

1次情報、本日のメインですが、われわれはテキストと言わないで「本文（ほんもん）」という言葉を使っております。後で紹介します。

2次情報では、文献資料の目録があります。古典籍総合目録とは、日本人の著した全著作の目録です。たいへん膨大なものになります。また、当館が所蔵している資料の目録があります。マイクロフィルムで持っている資料、あるいは少ないながら現物でも購入しておりますので、原本の目録などです。

それから、研究情報があります。例えば、雑誌に関する目録、あるいは論文そのものに関する目録などです。

その他のデータベースとして、国文学の分野で面倒なことは文字セットの問題です。古い文字をどうするかということは大変厄介な問題です。

また、用語データベースなども研究しています。さらに、著者や著名の典拠コントロールを行っています。同名異人や異名同人のシリーラスと言ってもよいものです。以下では、本文データベースに限定して説明します。

[OHP] (3) 古典テキストデータベース化のプロセス

これは、古典テキストのデータベースを作るプロセスの一例を表わしています。これを前提にして以下説明します。

まず、ソースという形で本がある。例えば源氏物語という本がありますが、これは、時代を経て現在のわれわれに伝わってきている。写本とか版本の形で伝播されてきた。今では、源氏物語という名前がついている本が、数百あります。それぞれ内容であるテキストが異なる。代表的なテキストでも8種類あると言われております。もしテキストデータベースを作って研究に役立てようとすると、極端に言えばその8種類のテキストが全部ないと駄目だということになる。本来、紫式部がどういうことを言おうとしたのかなどという研究はできないということになります。これは非常に膨大なものになります。

まず、一般的な流れで説明します。ある本を1つ選びます。それを仮に底本と呼びます。翻刻ということが通常行われます。これが写本ですと手書きの文字ですから、活字に直し、活字本という形で出版します。ここに当然本屋が絡んできます。さて翻刻は直接的に手書き文字を活字に1対1に変換する行為なのだろうか。それだけではなく翻刻者のある種の知的な作業の結果が入っているはずです。ある字をどう読むかということですから。

ここでは、便宜上1対1的な変換だと考えておき

ます。

さて、原著者、あるいは書写者、そしてこれを所有している人がいます。とくに、所蔵者の権利というのはたいへん重要です。図中に細かく書いてあるのは、いろいろな権利や義務の関わりをどのようにとらえていくかということのために関わる人という意味であげてあります。代表的なものだけを挙げてあります。

活字本では、これを作った人、すなわち翻刻者があるということです。それを出版するならば出版者、この場合は一応法人ではなく個人という立場で説明しておきます。出来上がりますと、これが流布されるわけですが、校訂本というものも出版されております。「校訂本」とは、さきほど、源氏物語には8種類ぐらいのテキストがあると言いましたが、どのようによむかなどということを校訂して、すなわち8種類の本をすべて校訂というプロセスを経て、1つの本、すなわち1つのテキストということに固定する作業と言えます。これを「校訂本」と呼びます。

これから紹介する岩波書店の『日本古典文学大系』というシリーズ本があり、これをデータベース化しようとしておりますが、これは校訂本です。従って、校訂本は、活字本とは違って相当な知的な所産物です。校訂者の永年の研究成果がここに蓄積されていると言ってもいいと思います。校訂本の校訂者はほとんどこの本を書いた人という意味で、著作者ということになります。当然出版社もここに関わってきます。

ここまでの段階では従来それほど問題はありませんでした。ところが、最近では少々やっかいな問題が出てきました。図中に“機械本”と書いてあります。これはあとで講演される内田保広先生が作られた言葉です。借用します。活字本をコンピュータに入力します。つまり機械可読化します。機械可読化するテキストの形態には2つあります。1つはプレーンテキストです。ワープロ的に、あるいはOCR的にテキストを加工することもなくそのまま入力するというやり方です。1つは、テキストは構造を持っていますから、構造を忠実に表現するという構造化テキストを作っていく方法です。いずれの場合も、機械本は1対1ではなくていろいろな変形が加わった1つの、活字本とは異なる1つの本が出来ることになります。今度は機械というシステムを抱える全体の所蔵者的な関わりが出てきます。

さて、機械本が出来上がった段階で、これはデータベースかというと必ずしもそうではなくて、データベースに仕立てることが必要になります。機械本を直接サービスするのではなく、データベースという形態においていろいろなサービスを行うことが必要です。そこで機械本をデータベース化します。機械本そのものは、機械可読化のテキストのデータですから、データベース化します。このとき、データベース化することによるデータベースの著作権が発生します。

また、データベースにはいろいろあります。例えば、われわれはオンラインデータベースを念頭においています。ところが、最近ではデータベースはパッケージ型のデータベースということがよく言われてくるようになりました。われわれの分野では「パーソナルユース」ということが大きな要件になっています。つまり、国文学者は、大体書斎に閉じこもって仕事をする研究態様ですから、大型コンピュータを使うよりも、パーソナルな環境で使えるシステムを望んでいます。もちろん電話ラインで使う場合もありますが、大体パソコンで仕事をすることが多い。そのようなところに情報資源を提供していく必要がありますから、パッケージ型のデータベースというのは大変重要と考えられます。

FDやCD-ROMなどのニューメディアで、提供して行くことになりますが、そのためにはそれぞれにやはり特有のデータ構造を作る必要があります。データベースから直接1対1に出力するということではなく、いろいろの変換や知的プロセスが関わります。その結果CD-ROMなどが生み出されることになります。したがって、ここにもデータベースの作成者が大事な要素として出てくることが分かります。

全体的な流れで見ますと、1つのテキストデータベースを作ってサービスすることには、これだけのステップを踏まなければならないということです。最近、これはとても面倒ですので、ソースから直接データベース化するという方法の実験を始めましたが、これは極めて大変です。

大変さは、校訂をどこでやるか、やはり図のようにプロセスを経ないと駄目かということに関わっています。標準のやり方はソースから翻刻を経て、かつ機械本を作って、機械本の中から、データベース化するということに落着かざるを得ないのかも知れ

ません。なお、ここに「印刷される」と書いてありますが、データベースから本を作るという方式も実験中です。

本文データベースを作成する側のプロセスということをまとめてみましたが、現実は必ずしもこのとおりに流れているとは限りません。国文学の方から見ると、「変だ」と思われるところがあると思われますが、このようなプロセスで、いろいろと複雑に多勢の人の関わりが出てきてますから、われわれはどう考え、処置して良いのか大変困っておるというのが実状です。

特に、最近の話題は校訂者です。校訂者の権利の扱いです。原著作者は既にいないわけですから問題ありませんが、校訂者は現在も活躍されている。しかも、校訂という知的生産物を作っている。従来は活字にする本を作るという意識で校訂という著作をした。ところが、これを電子媒体で出版するということについては、恐らく校訂者である著作者は意識してない。今後このあたりをどのように考えて行くかたいへん興味がある話題ではないかと思っております。

[OHP] (4) 本文データベース化の対象作品例

これは、現在本文データベース化の対象としている作品です。3作品挙げてありますが、この他にも幾つかあります。これらは、全部校訂本と考えています。

岩波書店が刊行した旧版『日本古典文学大系』があります。校訂本で、百巻あり、約600作品あります。600作品というのは日本の時代とジャンルをほとんど網羅するくらいの膨大な作品群です。古事記から始まり、江戸期のいろいろの例えば歌舞伎などの作品まで代表的な作品が網羅されています。

特徴は網羅性とういうことです。日本文学の流れとか文学史を研究する上で特に重要であるということで、データベース化の対象に選びました。ただし、一作品一テキストという意味で横並びという意識でとりあげています。文字数は約3,000万字、3,000万字と一言で言っても、データベースにすることの10陪以上の容量が必要になります。

それから、『噺本大系』があります。これは、東京堂出版社が2人の校訂者が校訂をした小噺約2万話の集成本です。20巻あります。これは岩波古典大系に対して、縦の方向に深くするつまり、ある特定のジャンルについてより多くの作品を集めてきて、それぞれの評価するために、選んだ対象作品です。これらは研究ユースということで、出版社から許諾を得てやっています。

[OHP] (5) 著作権にからむ問題のあれこれ

「問題のあれこれ」として著作権的に絡む問題ということで、テキストデータベース化の流れの中でまとめてみました。

まず、「多様なメディアによる提供」ということです。通常は、特定の作品の本文を冊子体で提供するということが主です。しかし、最近ではオンラインのデータベースで提供することの他に、パッケージ型メディアで提供するという状況が生まれてきました。問題点は、多様なメディアによる提供がいろいろな約束ごと、権利や義務の関係、などとどう関わってくるか、また、誰が処理するかという問題です。これは恐らくマルチメディアの著作権ということにもつながって行くのだろうと思います。

まず、基本的にデータそのものの著作権と、データベースそのものの著作権があります。さらに、そのデータベースだけでは無意味で、いろいろと活用するための機能ソフトウエアが必要かも知れません。そのとき、その機能システムについての著作権も必然的に発生します。それぞれ一ケ所でまとめてやっておれば良いのですが、恐らくバラバラになるでしょう。大変厄介な問題が出てくると思われます。

2番目に「多様なサービス形態」です。多様なサービス形態では、主としてサービスの主体、すなわち誰がどこでどのようにサービスをしていくかということで、主体をどのようにとらえていくかの問題です。もちろん、サービス品目毎の管理の問題もあります。資料や情報システムなどの情報資源の管理方法などの問題もあります。

3番目は、「多様なデータの存在形態」ということです。例えば研究用のデータの場合には、信頼性がないと役に立ちません。データの信頼性をどうやって保証するかということです。つまり、異本がかくさんあるから、その全ての異本を入力したら、ディレクトリサービスのようなことを起こして行かざるを得ない。恐らく新しい事業が発生するということになろうと思います。その新しい事業なるものをどのような形で起こし、運用して行くかという問題です。

4番目は、「横断利用」です。横断利用とは、いろいろなデータベースを渡り歩いて所望のデータを得

ることです。われわれは多数のデータベースを作ってきていますが、テキストを加えるのに本文データベースだけでは無理な場合があります。例えば本にはよく挿絵があります。挿絵を見るとその書かれている情景がよく分かります。言葉では説明しきれない情景描写は、絵に頼らざるを得ない。この場合挿絵とテキストをどうやってリンクさせていくかという問題です。

もう1つ例を挙げますと、今演能データベースを作ろうとしています。演能とは演じられる能のことです。これをデータベース化しようとしています。演ずる人々、演奏する人々、さらには関連して流派などがあります。また、使われる小物としての能面や衣装などがあります。さらに、謡や音曲などのような音楽があります。もちろんテキストは不可欠です。マルチメディア情報として取扱う必要があります。また、観客の反応といったような部分まで含めますと、そのようなデータベースをどのように作るかという問題があります。もちろん、作った暁には、どのようなサービスとするかの問題が生じてきます。

5番目は、「複製二次著作物」です。われわれは現在“分かち書き”ということをやっておりません。文章などの構造を表す標準のルール、先程SGMLの話がありましたが、われわれも同様のデータ記述文法を定めており、これに基づき標準化したテキストデータを作っております。これには分かち書きをとりあえずやっておりません。欧米の場合は単語に切れておりますから処理が簡単ですが、日本語の場合は単語に切れてない。これは大きな問題です。

われわれが、形態素解析などをやって語単位の確立を行うことはある程度可能ですが、やはり専門家の目から見るといろいろと問題があるようです。使う人がそれぞれの立場と環境で、分かち書きを行い、語の性質を定める必要があるでしょう。従って、最初は分かち書きをしないで、研究者からのそういうデータが集まってきたら、“誰それによる分かち書きデータベース”というのをまた作り、サービスして行く、そういうやり方をしようと考えております。

いずれにしましても、付加価値づけ、例えば品詞やいろいろの属性や索引をつける、あるいはコンコーダンスを作る、このような必要があります。索引で、人名索引例えば登場する人物、また地名などの索引を作ることが必要です。このような付加価値づけというのがどんどん広まってくると思われます。この場合には、その付加価値づけに関わる知的生産物としての権利もやはり考えておかなければならない。特に国文学の場合、これを制限してしまいますと研究はストップするということになりやすいから、ある程度自由に行えるような方策を考えなければならないと思います。

6番目は「品質のコントロール」です。例えば、著者ということ1つとっても同名異人や異名同人などの種々な問題があります。特に古い本ですと、誰が書いたかということ、すなわち、著者を同定しておく必要があります。この場合には、参照的なデータについて、多様なデータベースや印刷物などが必要になってきます。それらを利用して、つまりデータベースをいろいろと参照していくわけですから、多様なデータベースを渡り歩いて利用していく方式なりの実現という問題が発生してきます。

もう1つの品質コントロールでは、文字の問題があります。JISの制定が1978年ですから、コンピュータで日本語が扱えるようになってまだ10年そこそこということです。10年そこそこで日本語処理可能なコンピュータが常識となった。これはある意味では驚異です。

国文学研究資料館は今年で創立20周年になります。当初から、コンピュータで文字を扱おうとしていましたが、20年前のコンピュータはせいぜい片仮名が使えるという状況で、漢字を扱うことは大問題でした。

やっとJIS規格が出来、それにおんぶしてきているわけですが、古典の世界ではJIS規格の文字ではとても足りません。JIS外文字をどのように作るか、あるいは標準化して流通出来るようにするか、つまり、データーベースとしてどういう文字セットを設定しておけば良いか大変大きな問題です。

7番目は、「データ標準化」です。これはデータ記述文法です。われわれもKOKINルールと称するデータ記述文法を定めていますが、国際的にも活発な動きがあります。TEI (Text Encoding Initiative) です。そういうものを絶えずにらみながら標準のデータを作っていく必要があると思われます。

標準のルールを作ってデータを作った時に、それが固有の著作物と考えると、そのルールと著作物についての著作権が関わってきます。では、そのルールを使うためには何らかの許諾などが必要になってくるのでしょうか。いろいろな側面で関わりを持っ

てまいりました。

8番目は、「自由なデータ流通」です。1つの問題は、異機種間のコンピュータネットワークのためのプロトコルが必要です。それから、文字コードやフォントの標準化の問題があります。最後に、まとめとしていろいろな権利が関わってくると、一生懸命データを作ってもサービスできないということになりかねません。そこで、著作権などの取り扱いを考えてくれるような一種のセンター的なもの、それは組織であるべきか、あるいは委員会であるべきか分かりませんが、そのような状態が是非必要であろうと思っております。

それでは、以上で終わります。(拍手)

司会:ありがとうございました。それでは3番目に、内田先生、お願いします。