

漢字と情報

No. 3
2001・10



京都大学人文科学研究所 Documentation and Information Center for Chinese Studies (DICCS)
附属漢字情報研究センター Institute for Research in Humanities, Kyoto University



「尚書正義定本」と漢字コード

電子化とは何か？

ポスト文字コードの意義

人文研のアーカイブス(3) 『廣西通志』

「尚書正義定本」と漢字コード

木島史雄

本研究所が、まだ東方文化研究所であったころの研究成果の一つに「尚書正義定本」がある（第1冊 昭和14年 東方文化研究所経学文学研究室発行）。その名のとおり、中国の古典である『尚書』につけられた「尚書正義」という注釈の、現代に伝存する諸本を詳細に検討し、原本の旧に復すべくまとめられたものである。「この「定本」が「尚書正義」に対する最後の校定となることを希望し、またさう信じるものである」と校定者は記す（「尚書正義解題」東方学報 京都 第10冊）。たしかに氏の意気込みと自信に値するだけの価値を持ち、ある意味では近代日本における漢学研究の最も優れた成果の一つといってよい。しかし近年の研究環境の変化のおかげで、更にその研究を進めうる可能性がでてきたように思われる。

はなしは15世紀ヨーロッパへととぶ。1455年ころ、マインツでヨハン・ゲーテンベルグが金属活字による活版印刷術を發明した。そして印刷という手法ならびに現象がヨーロッパ世界にひろがっていった。ところで目を東洋に向けてみると、中国では明王朝の中ごろにあたり当時、印刷出版はごくありふれた活動、そして産業でもあった。道教のお経全集である「道蔵」1431種5305巻という大部の出版もこのころなされている。つまりこの時期、洋の西と東で、印刷物へのなじみ具合は大きく異なっていた。そしてそれぞれの世界に存在する印刷物の絶対量は、相当に大きく異なっていたということが出来る。

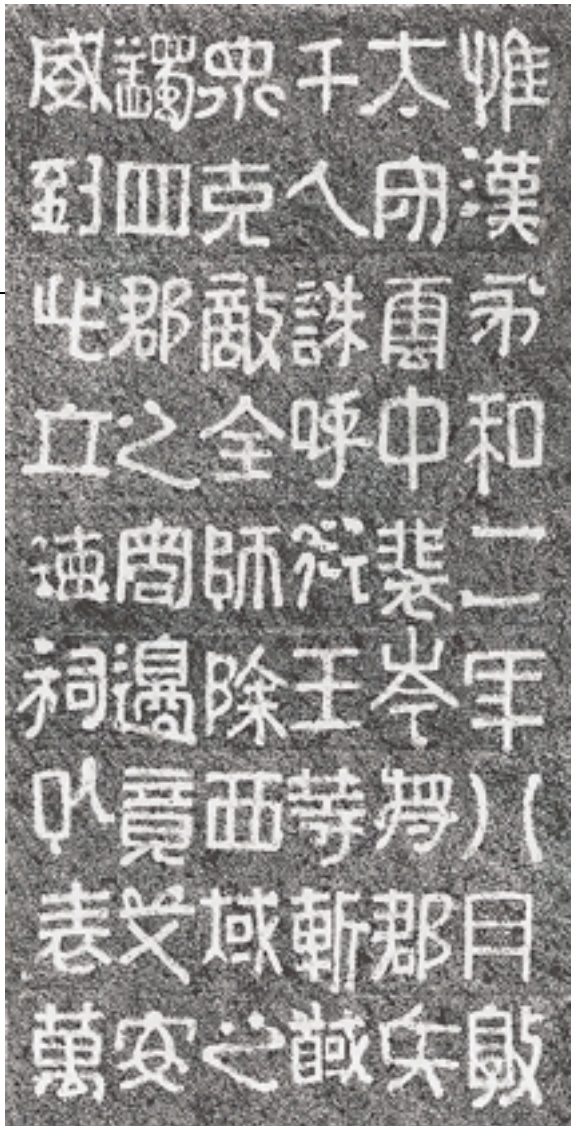
ところで一般に、印刷という技術が文化史に大きな影響を持ったその主要な原因は、書物の供給量にあるとされる。印刷のせいで、多くの人がキリスト教の「聖書」を目にすることとなったというわけである。しかし東洋における印刷の普及時期と生産量を考えれば、知識界への影響が、東西

で異なってもおかしく無い。では東洋におけるその影響とはどのようなものであったろうか？

東洋では近世以後、知識人はテキストを「印刷文字」として読むことに慣れきっていた。そして手書き文字が背負う字体の「ゆれ」やそこにひそむ事象への関心が低かったように見える。学術的関心も、印刷によって固定され、絶対性を賦与された文字の語義研究へ、もっぱら向けられた。すなわち書物の研究は、印刷された文字の連なりを出発点とし、先だつ鈔本時代の揺れ動く文字の姿は学術的関心から忘れられていった。これはあたかも、仏典が漢訳されると、中国の仏教研究がほとんど原語に遡ることが無かったのと同様である。金石文字についても、金石文そのものを研究するのではなく、文献をどう補強するかという点に主眼があったと見える。つまり東洋では、いち早く印刷術が普及したせいで、ある時期以後、印刷という技術の及ぼした影響は、書物供給量の増加にではなく、印刷文字による字体の画一化に、その主要な点があると思われる。技芸作品としての書に読みがたい文字がこのころから多出するようになるという大変興味深い現象もあるが、文字字体への関心は、近世以降の中国知識人の中で総じて薄い。そしてさらにいえば、異体字や古体文字を扱わないことが、文人としてスマートなことと意識されていたふしもある。玩物喪志というわけである。片や西洋では、漢字の字体にあたる綴り字の画一化、正書法の確立は、ラプレー（16世紀前半）の時期には進んでおらず、パスカル（17世紀半ば）の時代を待たなければならなかった^(*)。

さて「尚書正義」に話を戻そう。「尚書正義」が注解の対象とした『尚書』という書物には、「古文」という常態ならざる文字体を多く含む伝統があった。それは伝承によれば、秦の始皇帝の焚書ならびに文字統一以前の字体が生き残ったも

*1) フランス語の場合、オランダからの印刷物の流入によって近代正書法が決まったという。ちなみに現在本研究所の欧文紀要の名称は、背文字では「ZINBUN」であるが、冊子の表紙には「ZINBVN」とある。UとVの分化以前の状況を銜ったのだろうか。



のとされる。そして「古文」で記されていることが、正統的なテキストの証でもあった。したがって「正義」が採用した『尚書』本文も古文字体であった。その姿は、敦煌や日本に残存している鈔本によって確認される。そしてもちろんこの「尚書正義」が編集された唐のはじめには、印刷術は普及しておらず、印刷による文字の統一、絶対化はまだ起こっていなかった。

さて「尚書正義定本」である。この書の目的は、本文を含まない「尚書正義」のテキスト確定にある。しかし「尚書正義定本」は現実には本文も並べ記している。「尚書正義」の原著作当時の姿を目指すならば、『尚書』本文は無いほうがよく、入れるならば古文字体を用いるのが筋であろう。しかしそうはなっていないのである。もちろん「定本」の編纂者たちがここに述べたような『尚書』にまつわる事態を認識していなかったわけではない。それどころか本研究所には、敦煌将来ならびに本邦残存の『尚書』旧鈔本の写真が、徹底

的に集められ、收藏されている。そして「読尚書注疏記」(東方学報 京都 第7冊～11冊)をみれば、多くの『尚書』旧鈔本を詳細に比較検討して、校勘記が作成されている。しかし「注疏記」の時点で、すでに古文字体は議論に姿を見せないし、「定本」中には、古文字体の情報は見えない。その最大の理由は、『尚書』を読もうとするものへの配慮であるに違いない。というのは現在この世に存在する『尚書』のほとんどは古文字体を含まないからである。しかし字体に全く言及しないのは、いささか不親切であろう。このように思いをめぐらせてくるとき、さきにのべたような、東洋の学者たちの印刷文字への長年の狎昵と、手書き文字への無関心に思っていた。印刷文字をあつかうかぎり、文字字体などというノイズに煩わされること無く、本文の論理展開や、表現の巧拙に意識を集中できる。そして、「尚書正義定本」の編纂者たちも、このような流れの中にあっただのではないか。そしてそれはいかにも知識人らしい態度であると感じられていたのかもしれない。なお「尚書正義定本」の作成作業が進められていたころ、同じく本研究所で小林信明氏が『尚書』の古文字体にもっぱら目を向ける研究をしておられたが、残念ながら直接的にはその成果は「尚書正義定本」に取り入れられなかったようである。

さてわれわれの課題である。いち早くは電子複写機、最近ではコンピュータという道具の進化を受けて、あらたな視点、あらたな手法でのテキスト研究が可能になっている。そして本誌の前2号で語られた、コード体系からの漢字の解放が実現すれば、それは、ここに記してきたような文字字体をも視野に入れたテキスト研究にとって大きな力になるに違いない。そして、印刷というメディアが絶対性を失いつつある現在こそ、印刷が学術にもたらした影響についても、いま少し視野ひろく、しかも精密に考察する格好の時期であろう。「尚書正義定本」の成果を今一歩進める環境は整っているといえるのではないだろうか。(人文科学研究所助手)

電子化とは何か？ クリスティアン・ウィッテルン

電子化とは、あるものをデジタル形式で使えるようにする過程である。デジタル形式のメディアにおいては、数値を割り当てることによってすべてが表わされる。割り当てる数値の範囲と間隔が電子化の結果の品質を決めることになる。文字の電子化の過程も基本的に「間隔の細かさ」（「標本化」、sampling）と「割り当てられる数値の範囲」（「量子化」quantification）という2つの変数に準拠して変わる。

テキストの電子化はあるページの具体的な画像データを抽出することだと考えられる。ここでは基本的には同様なモデルを用いることが可能である。すなわち対象となるテキストの電子化において、サンプルとなる文字を選び出すことは標本化に相当し、選び出された各文字に対して数値を対応させる文字符号は量子化に相当する。

この様なモデルでは紙の上で見られる文字から2段階の抽象化を行なっている。まず具体的な文字の形を数学的表現に変換する抽象化、そしてさらに字形の数学的表現を各文字に対応した値（符号）で表すような抽象化である。この第2段階の抽象化の過程においても、多くの情報が失われてしまうのは明かである。例えば、文字の大きさに関する情報も存在しないし、フォントファミリーや書体の情報も失われている。このモデルにおいては、文字そのものの値のみを記録する。具体的な文字の形状を表していた数学的表現さえ超えていると言ってもいいかもしれない。今論じているのは、文字の単なる「アイデア」にすぎず、それはその存在以外は他の具体的な属性を一切持たないからである。このモデルによるテキストの電子化において、文字はその標本化周波数と数値の範囲において使えるアルファベットに割り当てられたものであるということになる。

ラテン文字を使って著されたテキストは、この値を対応させるのが容易であるため、このモデル

がふさわしいが、漢字で書かれたテキストを電子化するのには不適切である。漢字においてはテキストの符号化という行為は、この用字系（script）の持つ性質によって非常に困難なものになっているのである。概念的に漢字という文字は、意義素（sememe）あるいは形態素（morpheme）を構成する基礎要素にすぎないが、ラテン文字と比較するならば、文字よりもむしろ単語（words）に対応すると考えられる。

書記系（writing system）として漢字が用いられるところでは、時代や地域、あるいは文化の違いにより、非常に多くの字体の違いがあるため、上述した抽象化において文字を構成する対象を決めることはしばしば困難である。図1はこのことについて実例を示している。



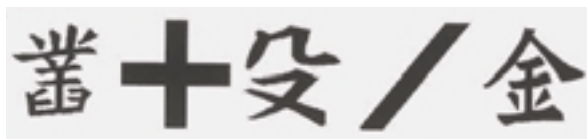
これらは高麗大蔵經に現れる文字（異体字）である。字形の違いは、時間や空間の違いによってもたらされるばかりではなく、ある少数の職人によって版木に彫られた、単一の文献にも現れることは注目すべきである。図1は李圭甲編『高麗大蔵經異体字典』（高麗大蔵經研究所刊、2000年12月。掲載した漢字はp. 1131ffのNo. 6535である）に現れる63体からの抜粋にすぎない。もちろんこれらはすべて同じ文字の異体字と考えられる（だからこそ編者はそれらを皆同じ文字として1つの項目にまとめたのである）。電子化の過程においてこの例に数値を割り当てる際には、それらをただ一つのコードにし、アルファベット書記体系で行うときと同じように、形状の違いを無視することができる。これが満足のいく解決法ではないことは言うまでもない。この方法は場合によっては

適切であるが（例えば情報の検索において、これはより高い再現能力を生み出すであろうが）、すべての状況に対して適切とは言えない。

ここで直面している根元的な問題は、漢字はラテン文字と違う言語学的な単位であり、これをそのままコード化することは英語の単語の1つ1つにコードポイントを割り当てるようなことであるため、ラテン文字と同じようにコード化することには無理がある。

そこで考えられる1つの解決策は、漢字そのままではなく漢字の構成単位（書記素, grapheme）をコード化することである。漢字は、誰の目にも明らかのように、いくつかの構成要素からできていることが多いのであり、それらを符号化するのはより簡単である。漢字の構成の規則性を考察した台湾中央研究院の謝清俊教授の研究によれば、2000余りの構成要素から最もよく使われる5万以上の漢字が作れるという。このシステムで表せない新しい漢字を作るために、新たなまだシステムにない構成要素が必要とされる可能性はほとんどゼロに近い。

図2は前掲の例としてあげた漢字が3つの構成要素からできていることを示したものである。



画面や紙に再現するためには、ラテン書記体系における合字についてなされるように、前もって作成しておいたグリフ（字形を抽象化したもの）から選び出す。作成されたグリフがない場合は、その時に作成することもできる。

以上ではテキストを構成する最も基本的で図形的な単位のことだけを考察してきた。しかし、テキストの電子化全般を論じるには、その他多くの特徴を考慮しなければならない。テキストは複数の段落から構成されているのであり、段落は文から、文は単語からできているのである。いくつかの段落が集まって節や章を作り、いくつかの章と

題扉、表表紙と裏表紙とで1冊の本となっている。

この構造を適正に処理するために、電子化にはより高次の抽象化が必要である。文書記述言語は根本的な構造のみならず、ある1つのテキストのその他多くの側面をもコード化するのに便利な方法として開発されている。

人文科学の様々な分野の学者が何百人も集まった国際的な共同プロジェクトである Text Encoding Initiative (TEI, テキスト電子化協会)は、“Guidelines for Electronic Text Encoding and Interchange (テキストの電子コード化と互換のための指針)”を編纂出版している。

TEI コンソーシアムによって定められたタグ集に加えて、コード化されたテキストを、単語やその意味についての情報を含むセマンティックなデータベースにリンクさせることは、多くの目的に有用である。英語については、プリンストン大学のジョージ A. ミラーが中心になって開発した WordNet 機械可読辞典に、コード化されたテキストをリンクさせ、セマンティック語彙索引を作ることによってこれに成功している。漢字の世界にも同様のデータベースを構築する必要がある。

本稿で概略を述べた電子化のプロセスは、抽象化の度を上げるプロセスとして組み立てられたものである。このようにコード化された情報は、目的に応じて、どのようなレベルにおいても使うことができる。しかし、抽象化することは細部が失われることを意味するのだから、コード化されたレベルでは答えられないような問題に際しての代替システムとして、詳細でより具体的なバージョンも保存しておくべきであり、またそうすることは可能である。これは活字版に正本のファクシミリをつけると同じことである。それ以外については、この新しい（抽象化された）バージョンは、コード化された情報を用いて再現することができ、画面や紙に印刷して示すことができる。こうしてデジタルテキストは現実に引き戻すことができるのである。（センター助教授）

ポスト文字コードの意義

師 茂樹

今は昔、妙安寺で写経の毎日を送る隠遁僧寿水の前に、毎夜、単（ひとえ）姿の女が姿を現すようになった。悲しげな顔をした女の顔には、あるはずの口がない。女が示した和歌を手がかりに、陰陽師安倍晴明はその女の正体を見破った。女の正体は、寿水が写経した般若心経の「如」の字だった。その字は、傍の「口」の部分に誤って墨を落としたために、染みになっていたのだった。

これは、現代の大人気コミック『陰陽師』の一話である。このようなモチーフは漢字文化圏の古典にしばしば見られるもので取り立てて珍しいものではなく、古典の現代的なアレンジとでも言うべきものであるが、コンピュータにおける文字処理の問題を考える上で、いろいろな示唆を与えてくれる。傍の「口」を墨で汚された文字の霊は、顔から口がなくなっただけで女の姿はそのままであった（安倍晴明が写経の染みを修復した後、顔に口が戻った女が現れ、晴明に謝意を表す）が、現代の一般的な文字コード処理では、漢字の一点一画が増減しただけで、何の関連もないまったく別のコードになってしまう。中国でよく用いられる点のある「器」と日本で用いられる点のない「器」との違いは点の有無に過ぎないのであるが、それ以外の共通点は一切無視されてその小さな差異だけで別のコードが付されることになる。

また、この文字の霊は、平安時代の女の格好をしている。中国が舞台であれば当時の中国女性の格好で、現代が舞台であれば現代女性の洋装で現れてくるのだろう。いずれにせよ文字は、地域や時代に応じて、換言すればコンテキストに応じてその装いを変えるものであるが、文字コードにそのような機能はない。中国の字と日本の字をまったく同じものとみなすか、まったく別のものとみなすか、という方策しかとられていなかったと言ってよいのである（CCCII や異体字タグのような試みはあったが）。

Unicode 3.1 が出るまでの Unicode は、共通化できる部分は共通化し、字形の定義に幅を持たせ、細かい字形の差異等、コンテキストに依存する部分の処理はユーザ側に委ねるといった“牧歌的な”仕様のおかげで、地域や時代を超える性質と超えない性質という漢字の両面性がかりうじて表現されていたように思う。ところが新しい Unicode 3.1 では、中国のしたたかな戦略によって牧歌的な有り様が崩れ、中国というコンテキストに親しい文字コードへと変質してしまった（詳細は川幡太一氏「新 ISO / IEC 10646 と Unicode の漢字を検証する」『漢字文献情報処理研究』第 2 号参照）。

守岡知彦氏が中心となって開発している UTF-2000 は、かつての Unicode のように厳密さを欠いた方法ではなく、計算機にふさわしい方法で漢字の多様な属性を表現しようという試みであると言えるだろう。現時点で各文字に付された属性は部首・画数情報、既存の文字コードとの対応等、コンピュータ処理において利用頻度の高いものが多く、またコンテキストに応じて振る舞いを変えるという点についても不十分な点がある（コンテキスト処理においては、コミュニケーション論や人工知能的なアプローチが必須と思われ、甚だ興味深い）。しかしながら、UTF-2000 のアプローチは、単なる背番号制にすぎなかった文字処理から文字本来の姿を取り戻す方法として、大いに期待される。

筆者も、Kanji Database Project (<http://kanji-database.sourceforge.net/>) という共同研究の場において、漢字の属性データを豊富にしデータベース処理に応用する研究を進めている。現時点では Unicode を中心とした文字コードにおける矛盾や混乱が自動的に炙り出されるようなデータが中心であるが、いずれは伝統的な属性古い時代の音韻や特定地域に特有の字体（国字など）の属性をコラボレーションを通じてデータベース化することで、漢籍データベースに厚み、深みを与える一助としたいと考えている。（早稲田大学非常勤講師）

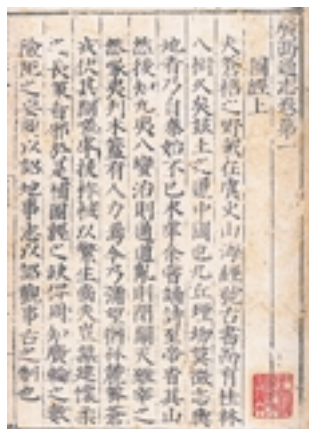
人文研のアーカイブス(3)

廣西通志 六十卷

闕卷第十六至第二十一第二十八第二十九

明林富修 明黄佐等纂

梶浦 晋



本所の漢籍目録の史部地理類では、一地方に関するものを〈地志〉といい、一定の体例を具備して、元以前に編纂されたものを〈古地志〉、明以降のものを〈今地志〉と呼び、目録では各々〈古地志之属〉と〈今地志之属〉に分類配列している。日本において今地志の収蔵にとむ機関としては、内閣文庫や東洋文庫などが知られているが、本所も必ずしも万全ではないが、少なからざる所蔵があり、原本で収蔵することがかなわないものは、景照本によって蒐集している。

ここにあげた『廣西通志』は現在の広西壮族自治区全域を対象にしたいわゆる〈省志〉である。現存する広西地方の省志は、明代から民国時期のものまで9種が知られている。このうち明代のものには嘉靖と万暦のものがあり、掲載書は嘉靖年間の編纂にかかるものである。嘉靖通志の今日伝存するものは少なく、国内では内閣文庫（藍印本）と本所の所蔵が知られるところである。中国においても稀で、中国国家図書館（北京図書館）と広西壮族自治区档案馆が知られる。

本書の大きさは30.0 (cm) × 17.8 (20.6 × 14.0)、金鑲玉装に改装。版式は四周単辺、白口、有界、10行20字。嘉靖壬辰（11年）蔣冕序、嘉靖辛卯（10年）林富序、および弘治六年周孟中序、弘治癸丑（6年）程廷珙序の二篇の旧序を付す。全六十巻を國經二巻、表八巻、志三十巻、列伝九巻、外志十一巻に分かつ。巻第十六から第二十一、第二十八第二十九の8巻を欠き、38冊に分冊されている。清末の文人で蔵書家としても名高い傅增湘の旧蔵で『藏園羣書經眼録』にも著録されるところのものである。ちなみに本所所蔵の（嘉靖）『廣東通志』（明談愷修、明黄佐等纂）もまた傅氏の旧蔵で、同じく著録がある。

印記は「翰林/院印」（左半に満文）「小山/堂書/畫印」「海虞科/第世家」（陰文）「蔣養菴/藏書記」「雙鑑/樓藏/書記」（陰文）「沅叔/審定」「増湘/私印」（陰文）「龍龕/精舍」「江安傅/沅叔攷/藏善本」「傅増湘/讀書」「書/潛」などがある。（センター助手）

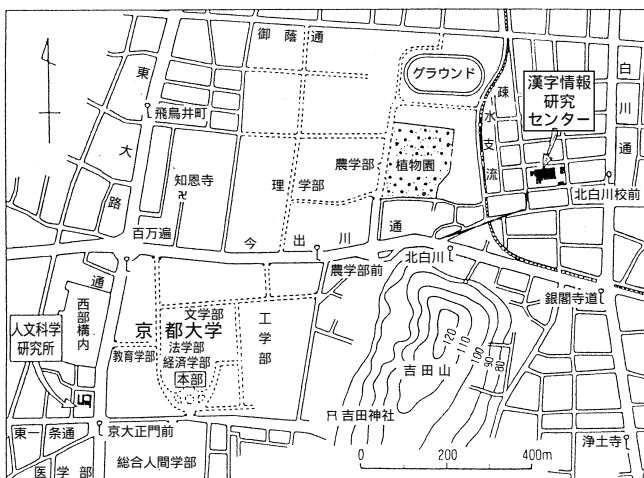
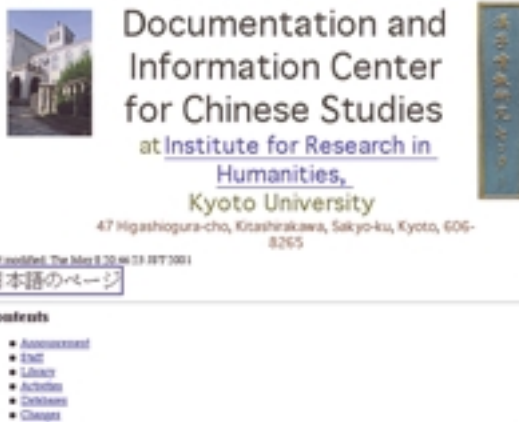
HP・TOPICS

漢字情報研究センターのホームページ (<http://www.kanji.zinbun.kyoto-u.ac.jp/>) では英語版を作成しました(下図参照)。公開中の CHINA 3 for WWW は、本年度末に新しい検索システムに切り替える予定です。利用者から著者名による検索もできるようにしてほしいという要望が多数寄せられていますが、新方式では可能になりますのでご期待ください。

京都大学人文科学研究所
 附属
漢字情報研究センター
 Documentation and Information Center for Chinese Studies (DICCS)
 Institute for Research in Humanities
 Kyoto University

〒606-8265 京都市左京区北白川東小倉町47

2000年4月、舊「東洋学文献センター」は新たな出発をしました。



【DICCS NEWS】

・10月1日(月)～10月5日(金)に第41回漢籍担当職員講習会(漢籍電算処理)を実施した。受講者は22名。本センターのスタッフに加えて、国立情報学研究所の宮澤彰教授、大型計算機センターの金澤正憲教授、小山田耕二助教授、沢田篤史助教授、高倉弘喜助教授、岩下武史助手、江原康生助手、川原稔助手を講師として迎えた。



・11月5日(月)～11月9日(金)に第42回漢籍担当職員講習会(初級)を実施する。参加予定者は28名。所外からの講師には、神戸大学の森紀子教授、立命館大学の上野隆三助教授、東京大学の橋本秀美助教授を予定している。

・11月19日(月)～11月20日(火)に平成13年度全国文献・情報センター人文社会科学学術情報セミナーが開催される。全体のテーマは「人文社会情報とIT」である。本センターが分担するセッションは、テーマが「漢字とIT」、発表者は安岡孝一センター助教授(「大漢和辞典とISO 10646」)、クリスティアン・ウィッテルン助教授(「漢字の電子化について」)である。

・最新のセンター刊行物

「東洋学文献類目」1998年度版(2001年3月) 井波陵一「中国目録学 四部分類法について」(2001年4月、講習会用テキストであるが、Web上でも閲覧できるように現在準備中)

発行日 2001年10月15日

発行所 京都大学人文科学研究所附属
 漢字情報研究センター

〒606-8265 京都市左京区北白川東小倉町47

電話 075-753-6997 FAX 075-753-6999

<http://www.kanji.zinbun.kyoto-u.ac.jp/>