KU-ORCAS

――オープン・プラットホームが切り拓く新しい人文知の未来

内田慶市

はじめに

KU-ORCAS（関西大学アジア・オープン・リサーチセンター）は、2017年度文部科学省私立大学研究ブランディング事業に採択されたもので、その目的は関西大学の特色ある豊富なリソースを基盤とする東アジア文化研究のデジタルアーカイブを構築し、その活用を通じて東アジア文化研究の世界的ハブ的研究拠点としてのブランドを確立することにある。

関西大学の東アジア研究は今から約250年前の江戸時代の「泊園書院」に遡るが、その泊園書院を源とする東西学術研究所（1951年創設）を中心に展開され、特に、2005年以降、文科省の学術フロンティア推進事業による「関西大学アジア文化交流センター（CSAC）」（2005–2009年）、私立大学戦略的基盤形成事業による「関西大学アジア文化研究センター（CSACII）」（2009–2013年）、更には2007年から2011年までの文科省グローバル COE の採択といった成果を挙げてきており、この KU-ORCAS はそうした研究成果の蓄積の上に打ち立てられたものである。

1．CSAC デジタルアーカイブの現状

さて、まず最初に CSAC 及び CSACII の研究プロジェクトで構築した「CSAC デジタルアーカイブ」について簡単に述べておく。

1.1　近代漢語文献データベース（2006年より）

このデータベースは2006年に筆者の科研費によって構築した近代漢語文献

資料のデータベースである。

特に19世紀の欧米人による漢語資料を中心に約400点ぐらいデジタル化して、全文検索、複数テキスト横断検索、簡繁体字検索、テキストと画像の連動表示を可能にしたものである。

ただ、すでに13年ほど前のもので、Mac だとブラウザに制限があるし、Flash を使うため大幅なバージョンアップが必要になっていたが、この原稿執筆時点でようやく新しいバージョンの公開が可能になった。

図1　近代漢語文献データベース（近日公開予定）

図2　旧近代漢語文献データベース

1.2　文献データベース

関西大学には下記のような東アジア関連の個人文庫が多数所蔵されている。このデジタル化にもこれまで鋭意取り組んできている。

内藤文庫（33500点）……内藤湖南（漢籍）
長澤文庫（30497点）……長澤規矩也（国漢籍）
中村文庫（33491点）……中村幸彦（国文）
増田文庫（16184点）……増田渉（魯迅、西学東漸）
吉田文庫　（2479点）……吉田伊三郎（アジア外交）
鬼洞文庫（10309点）……出口神暁（国文）
泊園文庫（16954点）……藤澤東畡・南岳・黄鵠・黄坡

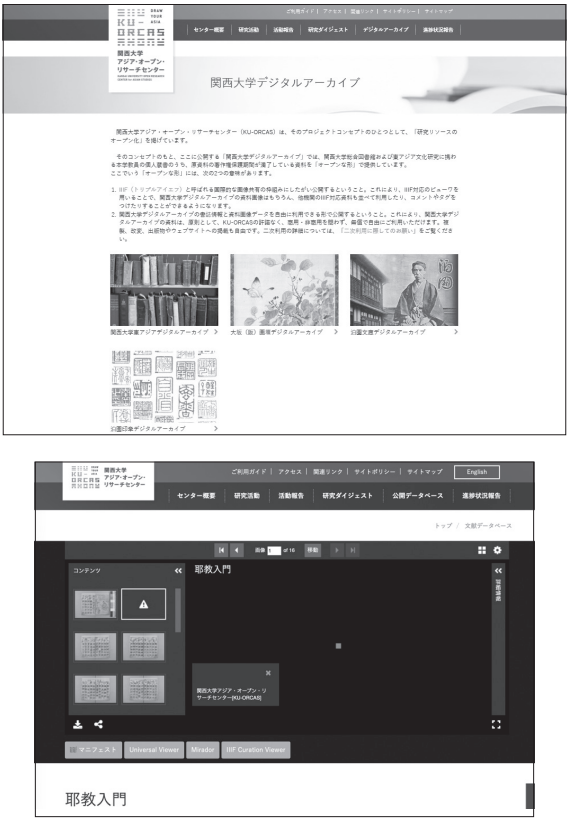

図３　関西大学デジタルアーカイブ（https://www.iiif.ku-orcas.kansai-u.ac.jp）

現在までに約6000冊程度のデジタル化が完了しているが、公開しているのはそのうち3000冊程度である。

なお、こうした個人文庫には書籍以外にも、書簡類、書画類や非文献資料も多数所蔵されており、そうした資料のデジタル化・公開も行っている。

例えば、以下のようなものがある。

漢封泥のデジタル化（20件程度）
内藤湖南の書簡が8000件程度：撮影終了
泊園関係の印鑑類が300点程度
内藤湖南文庫所蔵の軸物や貴重書庫収蔵の軸物700点程度（公開待ち）

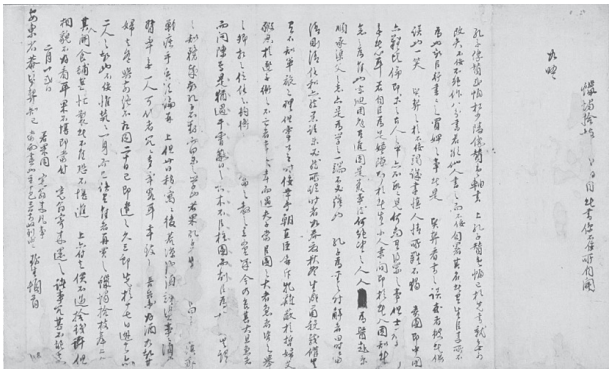

図４　朱舜水先生手簡（内藤文庫）

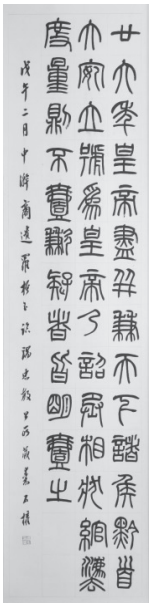

図５　羅叔言参事臨秦權條幅（内藤文庫）

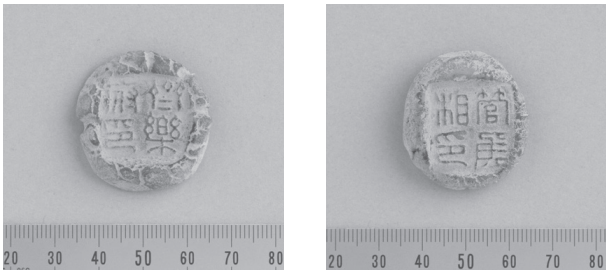

図６　漢代の封泥（Chinese-Style Wax Stamps）

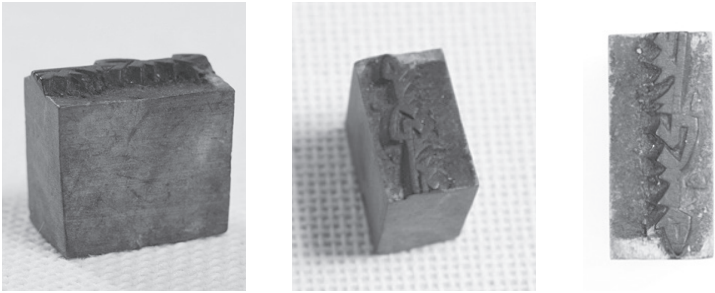

図７　ウィグル木活字（Uighur Print Stamps）
ペリオ (Paul Pelliot) から内藤湖南への寄贈品

図８　泊園印章デジタルアーカイブ

2．アジアにおけるデジタル化の現状

ところで、アジアにおけるデジタル化の現状であるが、最も早く（約20年前）から、しかも大規模にデジタル化を行ってきたのは、やはり台湾中央研究院の歴史語言研究所漢籍電子文献資料庫であろう。現在、約５億字のデジタル化が行われており、全文検索も可能である。

また、中国でも盛んに行われてきており、例えば、CADAL（China Academic Digital Associative Library＝中国数字図書館国際合作計画）は浙江大学と中国工程院による国家的プロジェクトとして2001年に開始され、すでに700万冊のデジタル化が完了している。ただ、日本ではこれまで国会図書館や私どもの関西大学も連携して加入していこうと考えてはきたが実現には至らなかった。最近ようやく東京大学が日本で初めて参画している。

この他、環太平洋デジタル図書館連合（PRRLA＝Pacific Rim Research Library Alliance）という国際的な組織もあり、香港、中国、台湾、マカオ、オーストラリア、カナダ、アメリカなど33の大学が参加しているが、これも残念ながら日本の大学は加盟していない（関西大学は筆者が図書館長を務めていた時代に一度加盟したが、現在は幽霊会員となっている状況である）。

日本に目を向けると、早稲田大学や国会図書館近代ライブラリー等は早くか

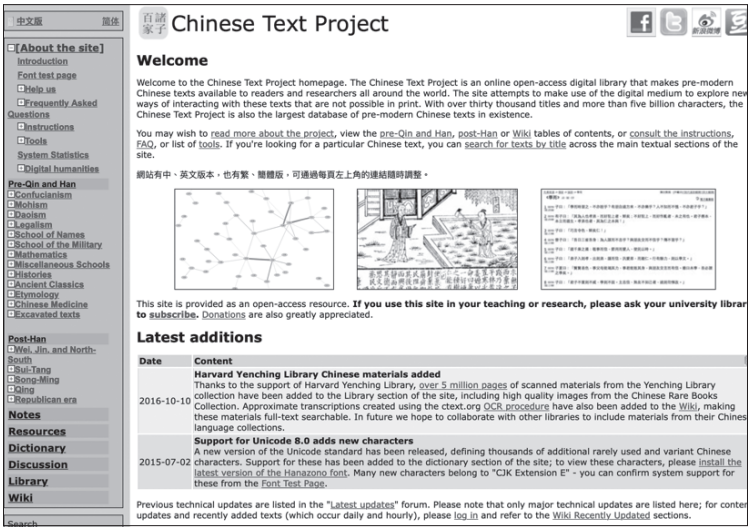

図９　CTEXT (Chinese Text Project) のページ（https://ctext.org）

らデジタル化を推進してきたが、ここに来て、国文学研究資料館の日本語の歴史的典籍データベース構築プロジェクトや京都大学、島根大学等々も積極的にデジタル化を推し進めてきている。

一方世界における東アジア文献資料のデジタル化の状況について簡単に触れておくと、例えば以下のような機関で多くの文献のデジタルアーカイブが実現している。

Gallica（BnF＝フランス国立図書館）
CrossAsia（ドイツ国家図書館）
Münchener DigitalisierungsZentrum Digitale Bibliothek（MDZ＝ミュンヘンデジタルセンター＝バイエルン州立図書館、IIIF 対応）
Digitalisierte Sammlungen（ベルリン州立図書館）
Digital Vatican Library（DVL＝DigiVatLib）
イエズス会文書館（Archivum Romanum Societatis Iesu）
Heidelberg University（COE）
Harvard-Yenching Library（ハーバード大学）
Serica（Bodleian Library, オックスフォード大学）
National Library of Australia（NLA＝オーストラリア国立図書館）

もちろん、Hathi Trust、Internet Achive、Google books といったいわゆるオープンアクセスサイトも充実してきている。

この他、中国語に特化したものとして以下のようなものを挙げておく。

中央研究院近代史研究所「英華字典」
CTEXT（中国哲学書電子化計画）
CCL（北京大学中国語言学中心）
BCC（大数据与語言教育研究所）

3. サイロ問題の打破

現在、国内外を問わず、多くの機関、組織でデジタル・アーカイブが進められてきており、デジタル化は今や世界の潮流であるが、ここに一つ大きな問題が存在する。それは「サイロ問題」と言われるものである。

つまり、多くの研究機関でそれぞれにデジタル・アーカイブス化が進められているが、横の連携が希薄で、各所で同じようなものが別々に作られて「閉じた」状態となっている。そして、自分たちのサイトに貴重な資料が沢山保管さ

図10　IIIF の実例（KU-ORCAS 所蔵の英華字典の対照比較）

れていても、閉じた状態であるため、アクセスも制限され、放っておく（アクセスしない）と腐ってしまう（アクセスが少ないからと維持されない）わけである。

こうしたサイロ問題が生まれる背景には、例えば、画像における国際的規格がないということがある。

こうした状況を打破するために、近年提唱されているのが、IIIF（トリプル・アイ・エフ＝International Image Interoperability Framework 国際的画像相互運用の枠組み）という試みである。

この IIIF とは国際的な Web コンテンツ共有の枠組みであり、国外ではスタンフォード大学図書館、英国図書館、フランス国立図書館、オックスフォード大学ボドリアン図書館等多くの図書館がすでにこれを採用しているが、日本でも東京大学大学院人文社会系研究科次世代人文学開発センター人文情報学拠点、京都大学図書館機構、関西大学アジア・オープン・リサーチセンター、国会図書館、国文学研究資料館等がこれを取り入れている。

この IIIF のメリットとしては、同じビューワー内に他機関の IIIF 対応資料画像を同時に見ることが可能であり、アノテーション（いわゆるメモ、注釈）機能を追加することも可能となり、資料の比較研究が容易となるほか、多様な立場の人の様々な解釈を共有できるようになることが挙げられる。今後、日本でも多くの機関がこの規格に準拠したデジタル化を行うことが望まれる。

4. デジタル・アーカイブと人文研究

上ですでに述べたとおり、デジタル・アーカイブス化は昨今の人文学研究の一つの大きな潮流となっている。それは学問研究に便宜を提供するだけでなく、実は人類の「知の遺産」の「保存」ということにもつながっている。またこうしたデジタル化、公開という流れはまさに筆者が以前から主張してきた「秘蔵は私蔵なり」という主旨にも完全に合致するものである。あらゆる文献資料は「公開」されるべきである、それが「書の使命」であるというのが筆者の基本的考え方である。

ただ、単にデジタル化、アーカイブス化だけでは不十分である。それは、研究者の立場から言えば、新しい研究方法と結びつく必要があるのだと考えている。

例えば、かつての語彙研究と言えば、カードの枚数に比例するものであり、ある語彙の初出は読んだ量に依拠するものであった。それが今やコーパス、コンコーダンス、全文検索、全語彙索引といったものによって研究方法は一変した。

また、こうしたツールを利用することで単なる語彙史研究から他の研究領域（概念史研究、思想史研究等）との領域を越境したコラボも実現してきている。様々な情報を追加していくことで、様々な角度からのアプローチが可能になるのである。

例えば、Google の Ngram Viewer を利用すれば、1500年頃から現在までのトレンドキーワードを調べられる。

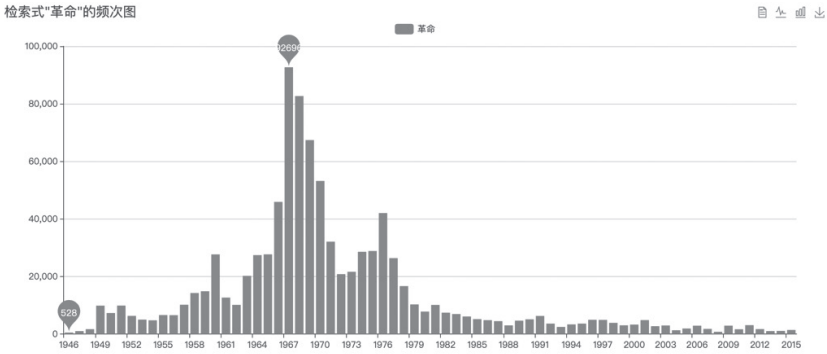

図11　Ngram Viewer での「革命」の検索結果

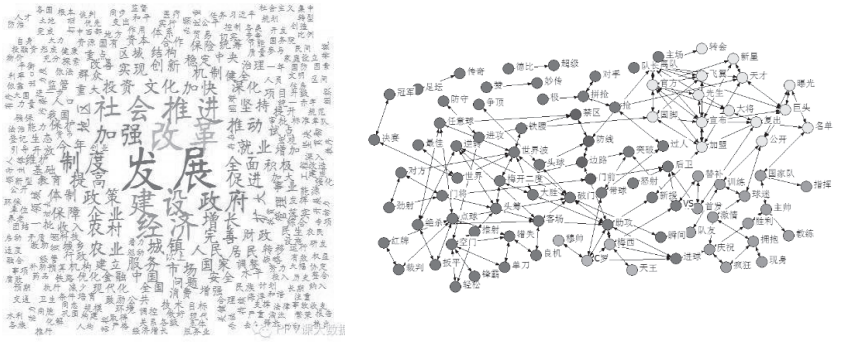

図12　中国語のテキストマイニングの例

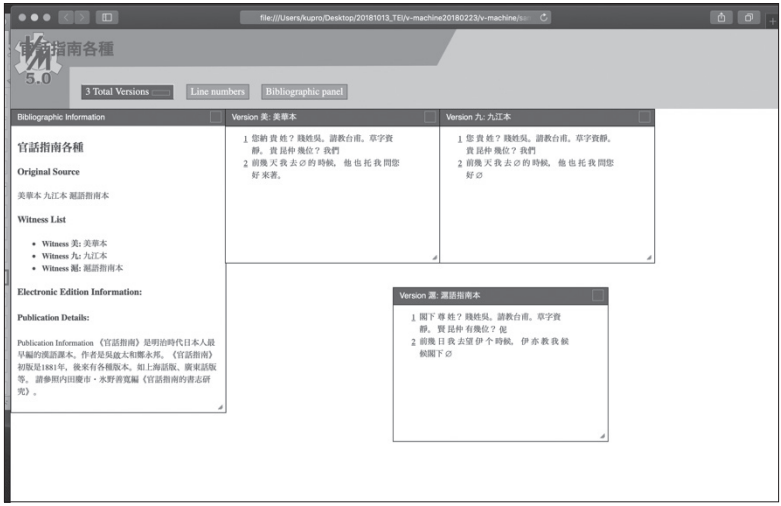

図13　TEI を利用した『官話指南』の校勘例

テキストマイニングでは、文章のデータを単語や文節で区切り、それらの出現の頻度や共出現の相関、出現傾向、時系列などを解析することで有用な情報を取り出すことが可能である。

この他、TEI（Text Encoding Initiative, 1987–）という人文学研究のための電子テキストの効果的効率的な共有のためのガイドラインあるいはその構造化のルールセットを利用することでテキスト（XML 化）は汎用性、永続性を持つことになり、例えば、テキスト校勘などを視覚的に行うことができる。

5．中国語とコンピュータの古くて新しい関係

中国（中国語）とコンピュータは実は極めて古くから関係がある。易の原理は陰陽の二元論であり、それはコンピュータの二進法に通じている。易の八卦は 3 ビット（三爻の組み合わせ）であり、上下で64通りの組み合わせ（すなわち 6 ビット）によって森羅万象を表現する。

ただし、古い関係にありながらも、いわゆる漢字や日本語はコンピュータに乗りにくい言語であった。私がコンピュータを始めた90年代でも「文華」等々のソフトがあったが、中国語の漢字を表示したり入力したりする場合は、第 2 水準の漢字ボードを中国語の漢字ボードに置き換えたりしていた。それが今ではユニコードの時代となり何万もの漢字を自由に扱うことができるようになっている。

それでも、中国語や日本語の自然処理は欧米語のように簡単にはいかないのだ。

今から約20年前に当時出入りしていたニフティサーブ（NIFTY-Serve）の FPRINT-15「せどおうくぱある（SED, AWK, PERL）」のフォーラムで知り合った彌永信美と齋藤希史に依頼して Perl と Apple Script で「全語彙索引作成プログラム」（1999.2.3）を作成したことがある。これは今でも利用できるし、現在は氷野善寛氏によりウェブ上でのプログラム（http://www.chlang.org/contents/index-converter/）が存在する。

使い方は、至って簡単で、まず以下のようなテキストを準備（文字コードは UTF-8）し、プログラムにかけるだけである。

［sample 1］
我 是 關西大學 的 學生 我 今=（これは次の行の単語と 1 語であるという印）
年 二十一 歳 我 學 漢語 専業 我
住在 大阪 我 有 父親 母親 弟弟
我 弟弟 今年 十七 歳 明年 要 考 大=
學 我 父親 今年 五十四 歳 母親 五=
十二歳

また、巻数、ページ数、行数を結果に表示させたいときには、先のテキストに次のように標識を付ける。

<V 1> ……巻数

<P 1a> ……ページ数（この例では、１葉の表。裏の場合は b で表す）
<L 1> ……本文の最初の行を 1 行と示す

結果は以下のようになる。

idx result of the file
/Volumes/MyDocument/Users/ni/Desktop/test_things/idx.pl/idex_pl_unicode_version/test_folder/chinese_test.txt

二十一 : 1 (1-1a-2)
五十二歳 : 1 (1-1a-6)
五十四 : 1 (1-1a-5)
今年 : 3 (1-1a-2, 1-1a-4, 1-1a-5)
住在 : 1 (1-1a-3)
十七 : 1 (1-1a-4)
大學 : 1 (1-1a-5)
大阪 : 1 (1-1a-3)
學 : 1 (1-1a-2)
學生 : 1 (1-1a-1)
専業 : 1 (1-1a-2)
弟弟 : 2 (1-1a-3, 1-1a-4)
我 : 7 (1-1a-1, 1-1a-1, 1-1a-2, 1-1a-2, 1-1a-3, 1-1a-4, 1-1a-5)
明年 : 1 (1-1a-4)
是 : 1 (1-1a-1)
有 : 1 (1-1a-3)
歳 : 3 (1-1a-2, 1-1a-4, 1-1a-5)
母親 : 2 (1-1a-3, 1-1a-5)
漢語 : 1 (1-1a-2)
父親 : 2 (1-1a-3, 1-1a-5)
的 : 1 (1-1a-1)
考 : 1 (1-1a-4)
要 : 1 (1-1a-4)
關西大學 : 1 (1-1a-1)
Total words : 24

ただ、このテキスト入力で最大の難点が「単語の区切り」である。中国語や日本語はこの単語をどこで切るかが極めて問題となるのだ。間にスペースを入れていくのだが、これまでは手入力で行ってきた。

最近ようやく、日本語でも中国語でもかなり優秀な単語を切るシステムが開発されてきているが、最終的にはやはり「人の手」が頼りである。ここがアルファベット言語との大きな違いである。

また、中国語の場合、Character（字）、Word（単語）、Phrase（句）、Sentence（文）をどう分けるかの問題もかなり厄介である。現在、中国国内でも以下のような「漢語分詞系統」があるが、現代語、近代語、古典語でも変わってくるし、やはり一筋縄ではいかないのであり、中国語学の専門家との「協働」も不可欠になっている。

THULAC　http://thulac.thunlp.org
HanLP　https://github.com/hankcs/HanLP
NLPIR　http://ictclas.nlpir.org

この他、中国語の場合、漢字の字体も問題となる。簡体字・繁体字・異体字をどう処理するかである。近年多くの電子ブックが世に出回っているが、テキストの信頼性に問題のあるものが少なくない。

例えば、「罷／吧」「很／狠」「里／裡／裏／裏」「您／儜」など、これらは違いがその成書年代とも関わってくるものであり、それを無視して一つの漢字で表記されたのでは、全く研究には使い物にならないのである。

6．多くの可能性——東アジア研究の Hub として

ところで、私たちの KU-ORCAS の研究ユニットとその主な内容は以下のとおりであるが、この他、広く学内外からの研究ユニットも募集している。

［ユニット 1 ：東西文化接触とテキスト］
本学所蔵の東西言語接触に関わる資料（辞書・文法書・宣教師報告書等）を中心としたアーカイブ。また、本学所蔵書のほか大英図書館・フランス国立図書館・バチカン図書館・ハーバード大学など海外諸機関の蔵書を相互リンクによって統合したものを構想する。

［ユニット 2 ：東アジアの中の大阪の学統とネットワーク］
本学の学統の源流たる「泊園書院」に関する総合アーカイブを構築する。

また、本学が集中的に所蔵する近世大坂画壇コレクションを中心に国内・海外に散在する大坂画壇作品を含めたデジタルアーカイブを構築する。

［ユニット 3 ：古都・史跡の時空間］

高松塚古墳の発掘に象徴される本学の古代飛鳥・難波津研究が蓄積してきた発掘データ・出土物データ・図面等をデータベース化するとともに、新たに飛鳥時代の墳墓の調査を行い、成果展覧会を開催し、これらを総合したアーカイブを構築する。京都の郊外都市・淀川流域の古文書・古地図・寺社境内絵図を調査・デジタル化する。

なお、こうした研究ユニットは決して「閉じた」ものではなく、私たちの KU-ORCAS の最大の特徴は上述の「サイロ問題の打破」を実現すべく、「コンテンツを解き放す」ことを目標にしていることである。

具体的には研究リソースのオープン化、研究グループのオープン化、研究ノウハウのオープン化であるが、全ての東アジア研究者（実は研究者に止まらず、一般市民、学生等々様々なステークフォルダーを念頭に置いている）に利用可能な研究プラットフォームを提供することである。

今後、様々な機関と連携しながら、東アジア研究オープン・プラットフォームの実現と知識基盤社会に適合した新たな人文知の構築に力を注いでいきたいと考えている。