Pioneering New Fields with Informatics Knowledge and Methods

The Trajectory of Research, Development, and Practice —Reflecting on a life in research

An Interview with

Professor Shoichiro Hara

to Commemorate his Retirement

An Interview with Professor Shoichiro Hara to Commemorate his Retirement

Pioneering New Fields with Informatics Knowledge and Methods

The Trajectory of Research, Development, and Practice —Reflecting on a life in research

Interviewer: Hiroki Baba
Program-Specific Assistant Professor,
Kyoto University Center for Southeast Asian Studies/Hakubi Center for Advanced Research

Leveraging an informatics approach spanning medical informatics, library informatics, humanities computing, and area informatics, Professor Shoichiro Hara has pioneered new academic fields and accumulated numerous achievements during his long career, whether at the National Institute of Informatics (formerly the National Center for Science Information Systems), the National Institute of Japanese Literature, or Kyoto University's Center for Integrated Area Studies and the Center for Southeast Asian Studies. We asked him to review his research and practice and to offer his perspective on the future of humanities–informatics collaboration as well as what is needed to advance collaboration between area studies and informatics.

——Before you became a researcher, what were your interests in high school and college?

My first interest as a high school student was biology. At the time (the early 1980s), there was much discussion around the subject we now call biotechnology.[1] Reading introductory books on genes and DNA that were just being published, an all-night experiment on sea urchin development during high school—these may have been triggers from which my interest in biology gradually grew. My rural high school also had a programmable calculator, literally an electric calculator that was about the size of today's desktop computers, and teachers allowed us to use it freely, which was truly a rare opportunity. In retrospect, I realize now that the calculator used a programming language close to assembly language.[2] It was not at all easy to use, but using it marked the first time I became absorbed in computers, as I was happy to calculate pi and the Tower of Hanoi.

I also enjoyed tinkering with machines, for example taking clocks apart, and building telescopes and radios. I read scientific articles and other publications about the imminent arrival of a society that would see the fusion of living organisms, computers, and machines. Perhaps because I grew up in such an environment, I became interested in learning about fields that mixed both biology and engineering. I decided to enroll in Natural Sciences II at the College of Arts and Sciences in the Junior Division within the University of Tokyo because of the biology and engineering programs offered there.

——After entering the University of Tokyo, you studied at the School of Health Sciences in the Faculty of Medicine, is that right?

At university, the fields at the "boundary" between biology and engineering available for study were

1 J. D. Watson was among those who received the Nobel Prize in 1962 for discovering the double-helix structure of DNA. In 1972, P. Berg et al. conducted the first-ever DNA recombination experiments. In 1976, W. Fiers et al. successfully sequenced the first complete genome. ↗

pharmaceutics, agricultural chemistry, and health sciences. Unfortunately, my score was a bit low for my first choice, so I decided to pursue my second choice, health sciences. My decision turned out to be a good thing. In the School of Health Sciences, I studied the basics of statistics, biology, medicine, and programming. I also conducted a field survey, although it was limited to Japan. This experience proved invaluable when I joined the Center for Southeast Asian Studies (hereafter CSEAS). So, all's well that ends well (laughs).

——When you were a student, I do not think many people enrolled in graduate school to become a researcher. Why did you decide to pursue master's and doctoral degrees?

It was because I was unsure of what to do after four years at university. So, I decided to postpone the decision and go on to graduate school. But I was pretty worried about entering the master's program at the School of Health Sciences. Epidemiology and biochemistry were my favorite subjects, and I got along well with the professors. However, health sciences had a strong sociological flavor, and I did not really fit in. Somehow, I also had poor relations with the professors there (laughs). While I was hesitating to continue with a health sciences major, I received an invitation to join the Institute of Medical Electronic Research Facility of the School of Medicine at the University of Tokyo. That led me to write my graduation thesis on the themes pursued at the Institute.

Still, I was not certain whether I should pursue a master's degree at the Institute of Medical Electronics. It had three divisions: Clinical Medicine, Basic Engineering, and Basic Medicine. The Clinical Medicine Division was researching artificial organs at the time. In 1980, they broke the world record for producing the longest surviving extracorporeal total artificial heart (in an experiment using a goat). The Basic Engineering Division was researching medical devices, such as electrocautery scalpels, and advanced tools such as biological sensors. The Basic Medicine Division, which invited me to join, was conducting medical data analysis and simulations and was researching non-invasive[3] biometric measurement methods. This matched my interests: although the professors in the division were physicians, the research they conducted was engineering oriented.

However, I thought that it would be a disservice not to study engineering first, so instead of going to the University of Tokyo, I entered the Department of Biomedical Engineering of the Institute of Medicine at the University of Tsukuba to complete my master's degree. As the Institute had an engineering laboratory, during my two years at Tsukuba, I studied the basics of circuits, control, and communication, and retook mathematics courses. I then returned to the Institute of Medical Electronics in the Faculty of Medicine at the University of Tokyo for my doctoral studies. During the doctoral program, I was asked to make a sensor circuit and a dynamic model of the circulatory system, so my two years at the University of Tsukuba proved to be quite helpful.

■The Beginning of Artificial Intelligence and Information Collaboration Research: Research on Diagnostic Expert Systems

——What kind of research did you do during your doctoral studies?

In those days (the mid-1980s), medical faculty professors strongly influenced research topic selection, and my supervisor assigned me to research an automatic diagnostic system for fluid therapy. My dissertation was on the same topic. Simply put, the purpose was to have a computer determine the optimal infusion volume and intravenous infusion rate to maintain normal levels of water and electrolytes[4] in the body. My research was to develop and write a program with three different algorithms. These were to estimate the degree of a patient's water and electrolyte deficiencies based on their symptoms, to create an infusion plan to compensate for the deficiencies, and to determine the rate at which the infusion would not burden the body.

I was immersing myself in this research during the middle of the second artificial intelligence boom. The first artificial intelligence boom had started in the 1950s. It focused on algorithmic research for inference and search, with the objective to intelligently perform games, such as chess, and to prove mathematical theorems. The second artificial intelligence boom began in the 1980s. It focused on incorporating human knowledge into computers to create intelligent systems that could be more flexible than those produced in the first artificial intelligence boom. Since these systems aimed to achieve the same inferential power as experts, they were called expert systems. The third artificial intelligence boom, ongoing since

2 String instructions corresponding to bit strings in machine language.

3 Non-invasive procedures are those that do not involve bleeding; ultrasound examinations, for example, fall into this category.

4 An electrolyte is a substance that ionizes into ions when dissolved in a solvent. Electrolytes in this study included sodium, potassium, and calcium.

2000, has led to the emergence of machine learning, in which machines themselves acquire knowledge by using big data. The third boom has also given rise to deep learning, in which machines themselves acquire the elements that define knowledge.

Returning to the topic, my research was to create an expert fluid therapy system. I began by reading thick books on fluid therapy and extracting the necessary knowledge. The first step was to manually convert the extracted knowledge into forms that computers could use. Next, I developed a database to store the converted knowledge and then wrote programs to execute inferences using the knowledge stored in the database. The mechanism for storing expert knowledge is called a knowledge base, and the mechanism for performing inference is called an inference engine. When combined, these two components form an expert system. During the second artificial intelligence boom, logical inference was the standard.[5] Therefore, the knowledge base comprised sets of logical formulas, and the inference engine was a "theorem prover." I dedicated four years of research to developing a theorem prover that could diagnose and prescribe infusion plans on the same level as an expert. Although the term "artificial intelligence" sounds glamorous, the actual work was an interminable and tedious process that involved reading technical books, rewriting knowledge as logical formulas, and upgrading the knowledge base. Creating the inference engine was also tedious, as it required rewriting and adding code every time a new type of knowledge was added, which rendered coding and version management painstaking.

I also had to verify whether the expert system I had built was usable. I would borrow medical records from specialist hospitals, read them, and extract symptoms. However, the old medical records were handwritten. Moreover, they were mainly in English, with some German mixed in. Deciphering the difficult-to-read characters to create patient data was akin to when Sugita Genpaku translated *Kaitai Shinsho* (*New Book of Anatomy*) (laughs).

When I fed my expert system with the patient data I had created in this way, the system would proceed to make inferences and output a prescription. I would then compare that output to medical specialist prescriptions as recorded in the medical records. Upon finding a significant difference between a doctor's and the expert system's prescriptions, I would first examine the expert system's inference process and explain the computer's inference process and result to the medical specialist(s). Next, I interviewed specialists about their inference and/or diagnostic processes and asked which part of the computer's inference process they thought was the source of the problem. I would then write the newly acquired knowledge into new logical formulas, which I added to the knowledge base. After the inference engine had been revised accordingly, the test would be repeated. However, the expert knowledge was sometimes vague, and some of their input was difficult to rewrite in a logical formula. Moreover, different specialists often had different knowledge, meaning their inference paths differed in many cases, although they produced the same prescription for the same symptoms. To reconcile differences, I continued to manually modify the knowledge base and adjust the program. When the modified program and knowledge were correctly adjusted, other parts that had worked well would go wrong, and I would have to readjust the system further.

In this way, it became clear that expert systems are flawed. Human beings must describe the knowledge, the high volume of knowledge makes processing difficult, contradictions among experts' knowledge are frequent, and ambiguous knowledge is difficult to convert to logical formulas. Other expert systems developed during my doctoral study days, such as automatic fault diagnosis of nuclear reactors and company complaint consultations, also suffered the same problems. Thus, the second artificial intelligence boom came to an end. Fortunately, just before the boom ended, I completed my doctoral thesis and received my doctorate.

Through my doctoral research, I came up with several ideas. First, since obtaining knowledge from medical specialists and organizing it without contradiction is difficult, the idea is to let the computer learn

5 If A, then B. If B, then C. Therefore, if A, then C.

on its own by using the information in medical records. This idea is the equivalent of today's machine learning using big data. I tried this after finishing my doctorate, but I gave up because medical records had not yet been digitized, which meant that the knowledge could not be automatically extracted. In addition, the learning program was practically unusable given the computing capabilities of the time. The second idea was to measure the amount of water and electrolytes in the body directly rather than inferring them. I abandoned this because I did not have time to develop measurement devices, and although the model was simple, I did not know the algorithms to calculate it. However, both these ideas led to my interest in medical information systems, which, in turn, led me to database and related research, which I have continued to pursue for the past thirty years or so.

As an aside, during my doctoral studies, I did not often use the term "artificial intelligence." This is because the definition or concept of "intelligence" was vague, and there was a lot of media hype about artificial intelligence. Instead, I called it computer decision-making or computer-supported decision-making. This was more in line with the research conducted in our laboratory. Looking back, I think I continued my research on computer decision-making, and even today, I do not use the term "artificial intelligence" often. Although the knowledge base changed from logical formulas to big data, and the composition of the inference engine changed from theorem provers to machine learning algorithms, the framework has remained the same. In recent years, my research focus has shifted from knowledge bases, which I studied for the past twenty-five years, to inference engines.

■Professional Experience at the National Center for Science Information Systems: Constructing a Full-Text Database System

——In 1990, you were assigned to the National Center for Science Information Systems, the predecessor of the present-day National Institute of Informatics. What did you work on there?

At that time, the primary mission of the National Center for Science Information Systems was to create a comprehensive catalog of the materials held by university libraries throughout Japan—in short, to consolidate the catalogs of university libraries. For those my age and older, it was impossible to know what kind of materials were available in other universities' libraries without actually visiting those libraries and physically perusing their catalog. Then, as university libraries across the country became networked, it became easy to search for materials that were not available at one's own university, in other libraries. Next, interlibrary loans allowed access to materials without visiting the holding university. The National Center for Science Information Systems promoted this initiative.

When I took on the new position, the catalog database primarily used a keyword search and only displayed bibliographic and location information for titles and authors of materials that matched the inputted keywords. Viewing material content was impossible, and the search would fail if the keywords did not match those that had been inputted. For example, materials with "computer" registered as a keyword would not appear in a search done with the word "calculator." But when the text of the materials is searchable, one can search using different words and see material content. This is a so-called full-text database, and I was involved in its development.

Incidentally, on a somewhat technical note, each material is comprised of text, such as its title and the author's name, followed by several chapters. A chapter begins with the title, followed by several sections. In other words, a full-text database needs descriptions of the text structure in addition to the character strings of the material. At that time, SGML[6] was established as an international standard for describing text structures, and I used it for text markup.[7] I did the markup manually.

Development of mass health examination system

6 Standard Generalized Markup Language, ISO 8879.

7 The process of adding structural information to the text, also known as tagging.

At the time, relational databases were the mainstream database systems used to store and search data (this is still the case today). Microsoft's ACCESS is a typical example of a relational database; intuitively, it is an image of a table. In such a table, each piece of data is itemized; for example, the field "Author" contains the author's name as its value. More technically, a value of a single item[8] must not have any structure.[9] This is called "first normal form." As I mentioned earlier, texts have structure, so a relational database cannot handle texts. Therefore, my task became to create a container for marked-up SGML texts and a mechanism to search them.

The search mechanism was the problematic part. Search languages such as SQL[10] were invented for relational databases and therefore could not be used for full-text searches. At the time, there was no search language for SGML text. To address this, a professor on my development team invented a search language called DQL.[11] However, he only wrote the specifications and told me, "You do the implementation" (laughs). What I did is too technical to describe here, but I developed a parser for SGML text. This parser analyzed SGML text and retrieved information, such as the text content with a specific tag[12] and the attributes of the tag.[13]

The DQL could write search instructions such as "Display the author of the document that contains the word ABC in the text tagged XX."[14] Therefore, I wrote programs that matched the DQL instructions with the analysis results of the SGML parser and output the strings within locations that satisfied specified conditions. The computer I used was a Unix workstation, which was becoming popular then, and the available language was C. I had studied databases and parsers after my doctorate, so that helped. However, I had no experience using a workstation or the C language, so when I think about it now, I am amazed that I could do that kind of work in two years (laughs). When I completed the system, I moved to the National Institute of Japanese Literature (hereafter NIJL).

■Contributing to the Work Being Done at the NIJL: Updating the Kanji Database System

——You moved from the National Center for Science Information Systems to the National Institute of Japanese Literature (NIJL) in 1991. I assume you were involved in similar database-related work, but did the content of your work change?

It changed dramatically. Instead of handling medical and natural science information and numerical data, I was suddenly dealing with classic literature and textual data, working with literary scholars as my colleagues. I have not always been good at Japanese, so when I came to the NIJL, I could barely read the classics, let alone interpret them (laughs). Furthermore, I had spent most of my education in science and mathematics classes and had almost no friends in the humanities throughout high school, university, and graduate school, so I had no idea about the temperament and research orientation of Japanese literature scholars.

——So, even though the general framework and direction of building a database were the same, the data you inputted and the people with whom you were working were completely different.

Exactly. It was tough to read classical characters (data), and although I had an interest in classical literary works, I was not interested in research on those works. At the same time, the features and processing of data are quite different in the humanities and in engineering. These were not insignificant problems, but from the standpoint of engineering, computers and databases are the same—this was the only thing that enabled me to work at the Institute of Japanese Literature.

My first task at the NIJL was to update the elec-

8 In Microsoft ACCESS, something that is written in a cell.
9 Such as repeating items, concatenating values, nesting, and so on.
10 Structured Query Language, ISO 9075, JIS X 3005.
11 Document Query Language.
12 E.g., content of the author tag is "Shoichiro Hara."
13 E.g., contents of the author tag is written in English.
14 E.g., the author of the document is "Shoichiro Hara."

tronic catalog database. The NIJL was founded in 1972 and is probably the oldest inter-university research institute in Japan. It had maintained an electronic catalog system since its establishment. I believe it was one of the world's most advanced humanities research institutes at the time. When I moved to the NIJL, the system was already nearly two decades old. As expected, the computer system was getting old and the software was no longer in line with the times, so my mission was to rebuild it.

In the early 1990s, servers at universities and research institutes were so-called mainframe computers, but Unix servers were gradually replacing them. Therefore, we had to decide to replace the mainframe computers as platforms.[15] Moreover, the computer network at the time differed for each vendor, whether it was NEC, Hitachi, or others. However, this was also the transition period to the Internet, so we had to replace networks and rewire the NIJL building to accommodate the new network. Webpages were also becoming popular, and we redesigned the user interfaces completely.

The problem was that the original database system was a "network database" older than the relational database. Moreover, although we had the original data specifications, as there were almost no records of subsequent modifications and updates, I had no idea as to the current state of the data structure. So, I began by extracting the data from the database as binary files written only with numbers. I then compared these byte-by-byte with the original specifications to find the areas where they differed from the specifications, or where there were numbers that were not in the specifications but that looked like data. When I found data areas that did not match the specifications, I manually converted the numbers into characters and asked a senior library staff member to identify the data. I repeated this process to create new data specifications. According to these specifications, I created salvage programs that extracted the correct binary data parts, converted them to text, and finally created SGML text. I used SGML to ensure data portability.[16] If the binary data remained, we would have to repeat the same conversion process in the next update, and I wanted to avoid this tedious job. After converting the data to SGML, I used my experience at the National Center for Science Information Systems to create a full-text database. It was a fairly advanced database for a humanities research institute in Japan.

——That must have been a daunting task.

Yes, it took more than a few years to complete. Another thing I struggled with at the time was Kanji. The Kanji code available for computers back then was comprised of approximately 10,000 characters spanning JIS Levels 1 to 4, but for Japanese history and classic texts, we believed that about 50,000 characters would be needed. Many of these were Kanji used for the names of people and places. Users had to define Kanji not included in JIS codes, which were called *gaiji* (external characters). NIJL's old database also used many *gaiji*, and I had difficulty importing them to the new computers. One of the problems stemmed from JIS code transitions. NIJL's old database used JIS78[17], the first Kanji code established in Japan. As this code changed to JIS83[18] and JIS90[19], some codes and symbols were replaced. As a result, if a computer was compatible with JIS 78, but a printer was compatible with JIS 83, characters displayed on a screen and printed on paper were different, even though they were the same Kanji (laughs). Finally, I wrote a program that traced all the changes in the JIS standards to correct these differences; the program contained instructions such as "I changed this JIS78 code to that JIS80 code." Around the year 2000, I changed codes from JIS to Unicode. Unicode includes most of the *gaiji*, so this problem was solved. The cumbersome task of Unicode conversions was determining the appropriate Unicode. As Unicode contains many characters with similar shapes, it is necessary to determine which code in Unicode corresponds to a specific *gaiji*. For this, I sought expert support.

15 A fundamental part of computing, such as operating systems and hardware.

16 Being able to easily migrate or move data stored on a specific platform or application to another platform or application

17 JIS C 6226-1978.

18 JIS C 6226-1983.

19 JIS X 0208-1990.

■Resource-Sharing Systems and a Full-Text Japanese Classical Literature Database

——You were affiliated with the NIJL for about fifteen years and involved in the Institute's research activities. What else did you work on?

Another primary task I worked on at the NIJL was resource sharing. Here, we aimed for an integrated search of databases distributed across the network. Standardization of data fields,[20] field names,[21] and data descriptions[22] were necessary to share information among databases. However, standardization is uninteresting for researchers who are taught to do something that has never been done before. Even if the data is just a bibliographic catalog, a researcher creates it with a particular purpose; that is, the structure of a bibliographic catalog differs for each researcher and research project. Therefore, some data omitted in one catalog may be described in detail in another. In extreme cases, there are as many different databases as there are researchers. In such a situation, integrating databases is advantageous for users, who find it difficult to search individual databases that adopt different descriptions and different data items. On the other hand, integrating databases can face a backlash from database creators, who are wedded to their different purposes and sense of ownership in creating the databases. Therefore, to facilitate resource sharing, we created a virtual database independent of each individual database. Data items in each database were mapped to data items in the virtual database. Users could search the individual databases indirectly via searches in the virtual database. This way, database creators did not have to recreate their databases, and users could search for data without knowing each database's location or data structure. I had this idea around 1999.

First, we tried sharing databases within the NIJL. We also tried sharing databases with Osaka City University because Professor Mamoru Shibayama, who was at the university at the time, was interested in the idea. It was around this same time that national universities and research institutes became corporations. The NIJL was incorporated into the National Institute for the Humanities (NIHU), and resource sharing became a project of the NIHU. The specifications and programs created under the NIJL's resource sharing project became the backbone of the NIHU project, and resource sharing was later developed at Kyoto University.

Another project of the NIJL was creating a full-text database comprised of 566 works in about 100 volumes of Iwanami Shoten's old edition of *Nihon Koten Bungaku Taikei* (*The Complete Collection of Japanese Classic Literature*). This project began before I joined the NIJL, and the data was almost completed by the time I arrived. The project still lacked a database and search tool, so my task was to develop these. Since the data were text data, I decided to use SGML. However, because this project started before SGML became popular, and the standard markup for humanities text, TEI,[23] was still being established, we used our own specific markup.[24] Shortly after I arrived at NIJL, TEI-P3 became available. However, as it had some difficulties handling Japanese classics, we decided to mirror the original tags created with KOKIN rules in SGML text, without using TEI.

The trouble with KOKIN rules, though, was their ambiguous definitions (laughs). Technically speaking, KOKIN rules should have been designed as context-free grammar,[25] but they were actually context-sensitive grammar.[26] Most programming languages, SGML and so on, are context-free grammar; standard parsers could not be applied to data written with different grammar, such as KOKIN rules. Even more troubling, the data had many errors, as they were manually created. These troubles meant that the original KOKIN data could not be easily converted to SGML text. To solve this, I wrote an error detection program to identify descriptions that did not conform to the specifications. I then fixed those manually. Next, I wrote a program to add supplement tags to the context-dependent parts of the KOKIN text and applied the program to KOKIN text to generate context-free KOKIN text. Finally, I wrote a parser to convert the context-free KOKIN text to SGML text. From this SGML text, we further created HTML[27] files for web

20 Different databases use different data structures, for example, in some, the title and subtitle are both inputs in the "Title" field, while in others, they are separated into "Title" and "Subtitle" fields.

21 E.g., a data item referring to a book title may be written differently, such as "Title" or "書名."

22 E.g., differences in writing dates, for instance, 30th January 2023 or 2023-01-30.

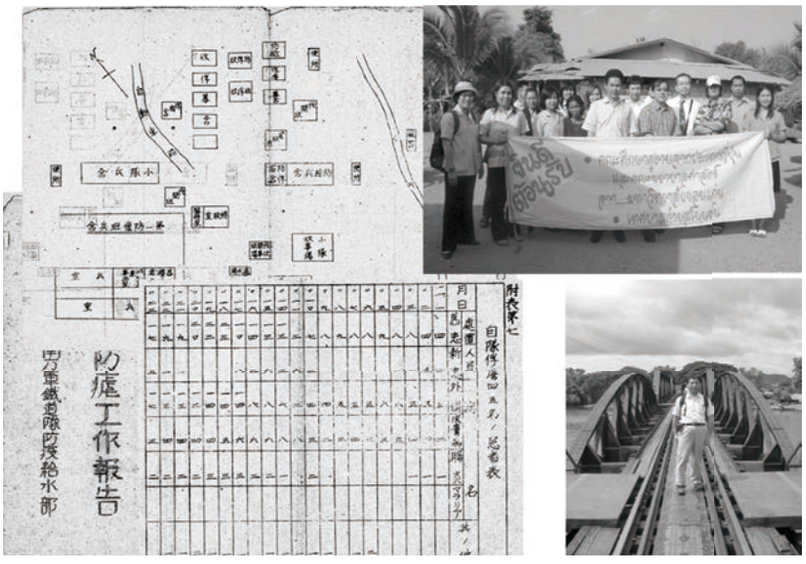
23 The Text Encoding Initiative.

24 Called the KOKIN rules ↗ (https://www.dlib.org/dlib/july97/japan/07hara.html).

25 Written formally, it is V → w. In other words, context-free means that the lexeme V can be replaced by another lexeme, w, without depending on the pre- and post-relationship with the lexeme V.

26 Written formally, αVβ → αγβ, where αVβ means the lexeme α and β before and after the lexeme V. In other words, context-sensitive means that the context before and after V determines whether lexeme V can be replaced by lexeme γ.

27 Hyper Text Markup Language, used to describe webpages on the ↗



display and PDF files for printing. I will not go into detail, but this part of the work required much time and effort. You can access the created data from the NIJL homepage.[28]

■Joining the Center for Integrated Area Studies (CIAS): Confusion Over Area Studies

——In 2006, you moved from Tokyo to Kyoto to join the newly established Kyoto University Center for Integrated Area Studies (CIAS). What kind of changes did you experience?

There were many changes. In addition to the research, the environment was significantly different. The starkest change was the climate. Fifteen years have passed since I moved to Kyoto, but I still cannot get used to the winter sky. Winter in Tokyo has sunny days with almost no clouds, but it is not so in Kyoto. Even when it does clear up, it often becomes overcast or snows soon after, and the winter landscape can be oppressive. I also found the place names in Kyoto difficult. For example, I do not know how to read "帷子ノ辻," but if I misread it, I may fear my surroundings (laughs). I was also very unfamiliar with the geography of the city. Nowadays, I can orient myself by looking at Mt. Hiei, but back then, I would get lost and confused by all the straight roads laid out in a grid and looking all the same. For a while after I moved here, whenever I came to the street level from the subway, I wasn't sure which direction I was facing (laughs).

My first confusion regarding the research concerned the term "area studies." I did not understand what it meant. Actually, this was also true for Japanese literature. Before I moved to the NIJL, I planned to buy a book on Japanese literature and study it, but I could not find any. I finally found a textbook from the Open University of Japan. However, it only contained articles on individual works, such as *The Tale of Genji*, rather than discussing Japanese literature more broadly. Similarly, I did not quite understand what area studies entailed. Indeed, I still do not understand it very well (laughs).

Perhaps because CIAS had just been established, the introductory phrase "My area studies are..." was often used when presenting research. This was a culture shock for me, as I had never heard anyone in the field of informatics say, "My informatics is..." even though informatics comprised various fields, for example, communications or data systems. I interpreted that they had to say so because researchers from various domains joined and had different methodologies. Then, on one occasion, while at a meeting, I said, "Area studies is not a domain, is it?" A senior professor yelled at me, "Yes, area studies *is* a domain!" I never said that again after that (laughs).

After all, I still do not understand many aspects of area studies. However, I see it as a field in which researchers from various disciplines study areas comprehensively; in other words, it is a container of multiple functions, like an aircraft carrier.

■Collaboration with Researchers from Thailand and Malaysia and with University Researchers

——Since moving to Kyoto, which activity or research project has left the most lasting impression on you? You must have participated in joint research and technical support.

Beside database construction, I had little experience with collaborative research at CIAS or CSEAS. However, I often collaborated with researchers in library and information science, health sciences, nursing, and humanities computing,[29] mainly from Thailand and Malaysia, and from Taiwan and the United States. It was very exciting to work with people who were conducting similar research in their respective regions. For example, in Thailand, I participated in a project to develop a database of health surveys of rural residents conducted in connection with community nursing activities. I helped design the database. This experience gave me pleasant memories of visit-

World Wide Web.

28 https://base1.nijl.ac.jp/~nkbthdb/

29 I do not distinguish between "humanities computing" and "digital humanities." I prefer to use the term "humanities computing," but because "digital humanities" is now popular, I use "digital humanities" here.

ing towns and villages in Thailand that I would not have been able to visit by myself (as I cannot speak Thai). The survey was a complete enumeration of each rural administrative unit, or *tambon*, which covered non-urban areas across Thailand. Hence, the data were fascinating from an academic point of view. One problem I experienced several times was that, perhaps due to our poor English skills, although we would reach a mutual agreement about data structure at a local meeting, after I returned to Japan and looked at the data from Thailand, I would be shocked and wonder, "Why did they send me this data?" (laughs).

For me, the research activity of the Unit for Academic Knowledge Integration Studies left a lasting impression. It is one of four units belonging to The Kyoto University Research Coordination Alliance, a coalition of research institutes and centers affiliated with Kyoto University that aims to create new academic fields across research areas and experiment with interdisciplinary research. It was established when I was the director of CIAS. I was in charge of the operations of the unit, which aimed to develop databases to facilitate advanced usage of Kyoto University's research materials as "academic knowledge" and to promote research using these databases.

Working as a leader of the unit gave me valuable experience. In addition, my appointment as director of CIAS gave me more opportunities to converse with professors from other departments of Kyoto University, which was very stimulating and impressive. This interaction has been a valuable asset in developing my research. Usually, I prefer to stay in my room reading books and writing programs. However, because I was required to attend regular meetings with department deans, I had more opportunities to exchange ideas with professors from, for instance, the graduate schools of agriculture, medicine, and energy science. This enabled me to encounter exciting ideas and make new connections. Since I specialize in informatics, my relationships with the professors in the Academic Center for Computing and Media Studies in particular grew more robust, leading to the big data research I am currently conducting.[30]

■The "MyDatabase" System for Area Studies Scholars: Simplifying Database Creation and Use

——At CIAS, the MyDatabase system was developed and implemented to allow area studies researchers to build and use their own databases. Could you tell us more about it?

As I mentioned when discussing the NIJL database, diversity is essential to research databases. This point is no different in the realm of area studies. Of course, from the perspective of the database system manager, it would be straightforward if, for example, I could create a single catalog database system, and everyone would input their data there. However, since that is impossible, CIAS's database system—called MyDatabase—had to deal with data diversity. This created challenges in both the data creation and data usage functions of MyDatabase.

On the creation side, data creators do not have enough data literacy, so to speak, regarding what requirements the data must meet before a database system can accept it. When we began developing the database system at CIAS, many colleagues provided me with a great deal of data. However, none of these data could be directly converted to a database because the data did not meet the requirements of a database. This may have been because some researchers may think that Excel files are enough for databases, which is not the case. For example, a data table may include other tables (a nested table), or the same item[31] may appear more than once in a table. As mentioned earlier, these data structures violate the "first normal form." These structures may be convenient for humans to read, but database systems cannot handle them. Other examples include non-standardized descriptions of data item names, such as describing a data item name over two lines (to represent an item name with explanations together) or inserting spaces in a data name (to write a data item name as an English sentence).[32] Such descriptions are not allowed in many database systems. There were many other irregular cases, such as the mixing of numbers and characters,[33] or different data types.[34] In the end, I had to manually modify each or write correction programs for each data file. What bothered me was that, even if I solved a problem for one database, another researcher would bring data having similar problems. Therefore, while it is impossible to handle all exceptions, I should have designed the CIAS database system to be flexible to some excep-

30 "Efforts in the New Field of Area Informatics Using AI and Big Data" https://onlinemovie.cseas.kyoto-u.ac.jp/en/movie_hara/

31 For example, ten fields called "Author Name" may be prepared in advance.

32 For example, "Title of Authors," where something like "Title_of_Authors" or "TitleofAuthors" would be better.

33 E.g., the integer 1 is not the same as the character 1, and in strings, 01 is different from 1.

34 E.g., the integer 1 is different from the real number 1.0.

tions. Specifically, the CIAS database system can handle data with repeated item names, spaces in the item name, and has no ID requirement. MyDatabase allows data to be registered into the database if it meets certain criteria, even if it does not entirely meet the strict requirements. I will skip the explanation, but another prominent feature of MyDatabase is that no data definition statement is required.

As for usage function, researchers are very particular about how they use data, or rather, how they use data according to the purpose of their research. Even I would change my approach according to whether I was wearing my information service provider or researcher hat. As a researcher, I change the user interface often. In the past, information systems required a fair amount of programming to create and maintain user interfaces. Fortunately, now that the web homepage is the foundation of the user interface, it has become relatively easy to create user views. Anyone with knowledge of HTML can write this part.

On the other hand, the database system operation remains as complicated as ever. Therefore, I have limited the database system functions available to users and defined procedures necessary to use these functions. These procedures are called an Application Programming Interface, or API. MyDatabase has a dedicated API. Thus, users can receive necessary bibliographic and image data from MyDatabase by composing simple search instructions according to the API and sending them to MyDatabase. All that remains to do is to create a homepage using the received data and complete the desired user interface. Although API is a popular function today, it was a novel feature just a decade ago.

■New Possibilities Due to Informatics' Permeation of the Humanities

——The third artificial intelligence boom is gaining momentum. What are the possibilities for fusion between the humanities and informatics?

I do not know if we should call it "fusion," but I think the synergy effect will give rise to new research fields. For example, when I was in graduate school, the development of CT (computed tomography), high-performance sensors, and medical expert systems during the second artificial intelligence boom led to the establishment of the Japan Association for Medical Informatics. Now, medical informatics is a medical research field. Corresponding to Medical Informatics in the field of medicine, in the humanities, we have digital humanities, the earliest example of which wan can trace back to 1946.[35] Today, entities like the international Alliance of Digital Humanities Organizations and related international journals have been formed. In this sense, I think a new "fusion" between the humanities and informatics is possible.

Although not an application of artificial intelligence or machine learning, there are examples (in the field of Japanese classical literature) of computers making it possible to conduct analyses previously considered too difficult. For example, the technique of *honkadori* in

Database development in the Community Nursing Project in Thailand

35 When Roberto Busa created Index Thomisticus (https://www.corpusthomisticum.org/it/index.age).

waka poetry incorporates a portion of an old poem into a new poem. Introducing *honkadori* is intended to create more complex and richer expressions and identifying a part of *honkadori* seems to be a research subject in Japanese classical literature. One group of researchers describes applying pattern-matching techniques from bioinformatics and other fields to analyze *waka* strings and reveal parts of original poems.[36] In another case study,[37] a computer is used to statistically analyze the characteristics of lexical occurrences to authenticate classical texts. We can certainly expect to see an increase in such cases.

However, I think a little more contemplation about whether we really need computers for humanities research is in order. Even if computers are useful for organizing resources, computers may not be necessary if the research is mainly about interpretation. In fact, I sometimes think that interpretation is at the core of humanities research. If that is true, digital humanities would be superficial humanities research. However, humanities research may radically change if computers can write literary works. This is very interesting. So, I think that the latest information technology, artificial intelligence and so on, will become widespread among those humanities researchers who say, "With the help of computers, I can do things that were previously impossible," or "Without computers, my research will not progress." Personally, however, I am a bit doubtful about what will become of the humanities as a whole in such circumstances.

——Recently, a smartphone app was developed to analyze and read corrupted Japanese characters. Even if researchers do not adopt this app, it seems it will spread to the general public.

That is possible. I say that because if you look around, from the perspective of information *technology* rather than information *science*, you will find quite a few tools and applications that have permeated the general public, even if they have not been accepted among researchers. A typical example is machine translation. I like science fiction, but it is not easy for me to read original texts. When I just want to follow the plot, I would be happy if a machine could translate a text to Japanese, even though there are many mistranslations. This situation is not limited to machine translation; the use of technology may eventually also affect the study of information "science."

Again, I digress, but that phrase "even if they are not accepted among researchers" reminds me of something. In 1999, I worked as an overseas researcher at the University of California Berkeley. The university IT team developed an e-learning tool for language instruction and asked the language faculty members to enter teaching material data. The person in charge of the project told me that the faculty teaching Hindi, Arabic, and Chinese were willing to provide input, but the Japanese faculty were reluctant. I thought, "Why wouldn't they adopt it?"

I conjectured whether this reflected something peculiar to the Japanese vis-à-vis new information

36 https://www.ism.ac.jp/editsec/toukei/pdf/48-2-289.pdf

37 https://www.jstage.jst.go.jp/article/sicejl1962/39/3/39_3_216/_pdf

tools. Perhaps they considered e-learning to be somebody else's domain, they wanted to avoid challenges and continue to do things their own way, or they were afraid to use the tool due to low information literacy. Considering how young Japanese today use smartphone applications for everything, it is a bit hard to imagine.

■Information Literacy to Support Informatics-Area Studies Collaboration

——Regarding research, what issues do you foresee in advancing future collaboration between informatics, area studies, and the humanities?

When informatics researchers participate in humanities research institutes like CSEAS, they face difficulties in achieving research results. This is especially true because informatics researchers often end up assisting their colleagues. For example, when building a database, we allocate considerable time and effort to building tools and manually collecting and analyzing data. Unfortunately, despite the considerable time and effort involved, the result has low originality as informatics research, making it difficult to publish papers and in turn, receive promotions. In my case, I was lucky because during my time at NIJL and CIAS, few researchers were engaged in digital humanities and creating databases was valued in its own right—but this is not true today. Young informatics researchers coming to humanities institutes must conduct original informatics "research" while also creating databases and managing information systems. Therefore, I think it will be essential for informatics researchers to try to position informatics as a research area rather than as support for area studies. To do this, area studies researchers need to try to increase their literacy as well.

It is not easy to have a conversation without literacy, making collaboration challenging. For example, during database creation, those on the information "side" must gain a certain level of literacy of the client side before starting the work. In my case, when I was a member of the NIJL, I studied the necessary vocabulary used by the scholars and librarians there and then analyzed the differences in organizing classical and contemporary materials before attending preparatory meetings. However, the scholars and librarians often did not make the same effort. Moreover, they sometimes used information technology terms, but, in many cases, they had not looked up the correct meaning of the terms, opting to interpret them as they saw fit instead, which sometimes became a source of confusion during discussions.

——Younger students have better basic information literacy; for example, they can all use Excel, so in that sense, things may have improved a little.

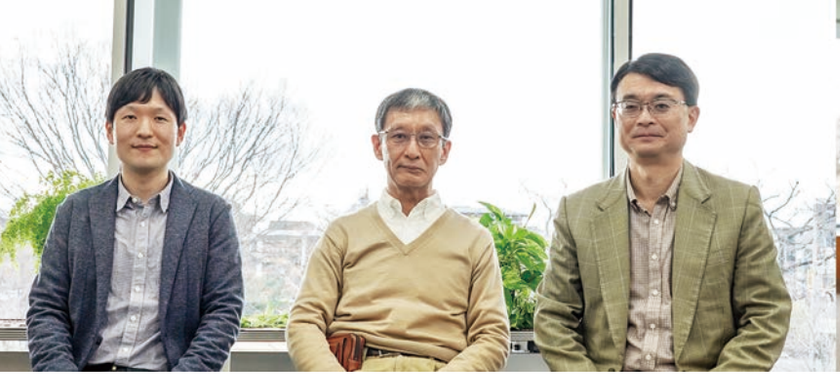
If we are only discussing databases, people who can work with Excel are, on the contrary, somewhat troublesome. It is like putting an image on a table (laughs). Knowing such techniques may be better than nothing, but it is not information or data literacy. At a minimum, humanities researchers should know what a database is and understand the first normal form. Some researchers cannot distinguish between anomalies, outliers, and missing values. This may seem trivial, but the difference is essential. Also, if one is going to do statistical analysis, they should at least have introductory-level statistical knowledge. Nowadays, we have convenient tools available for free, so obtaining some analytical results is simple if you have data. However, I have seen some cases where a researcher uses a tool or algorithm regardless of whether it meets the data requirements it assumes.

While all this may be true, it is unreasonable to ask humanities researchers to study informatics thoroughly. It is also challenging for them to devote a significant amount of time to creating perfect data. On the other hand, we in the informatics field do not want to spend most of our limited research time helping others. Therefore, creating a position to support data science is necessary. Although funding may be an issue, I think there should be at least one person in every organization who specializes in data science, such as a URA (University Research Administrator) or a research assistant.

■Advice for the Next Generation of Researchers: Be Ready with 2+1 Topics

——Do you have any advice for researchers who want to get into informatics or related academic fields?

I have occasionally heard people in data science say, "I am a data analyst, so I do not know the detail about the data itself." This is very curious. I question researchers who present results based on poor-qual-



ity data or without reviewing the data's validity. Although algorithms are essential, it is not good for researchers to use computers if they have not acquired the ability to read data or an adequate level of data literacy.

I do not want to sound old by giving advice, per se. However, I will repeat something that my supervisor told me, which is that if you are involved in research, you should have two main themes and one sub-theme, for a total of three. I want to share this advice with others, because if you are devoted to only one topic, there will inevitably come a time when you get stuck. But if you have two research topics, one of them will always be in motion. Many researchers have studied the same topic in humanities for a long time. However, in engineering or natural science, trends go out of style quickly, so you must be prepared for your next move in case one topic is gone in a few years. To be ready for such a scenario, you should keep a sub-theme in mind as well.

I did not follow my supervisor's advice faithfully, but I have found myself conducting research mostly by adhering to the 2+1 rule. At the NIJL, my main work concerned text databases and resource sharing, but I also started research on the application of GISs (Geographic Information Systems) to the humanities. By the time I moved to Kyoto University, this had become my primary research topic, which led to the development of spatiotemporal processing tools and the creation and publication of a historical gazetteer.[38] Machine learning—which began as a sub-theme—became my main research focus during my work at the Unit of Academic Knowledge Integration Studies.

In science and technology-related research, it is possible to branch out from a single research topic (my trajectory was artificial intelligence → databases → big data). Therefore, while it is good to have a single base, you should have 2+1 themes in mind.

——Indeed, if one persists with single-themed research, it is sometimes difficult to make progress, isn't it?

Exactly. I remember being upset when I first heard this advice. I thought, "I don't know if I can do that" (laughs), but now that I think about it, it was wise advice.

■Source of Ideas and Inspiration: Science Fiction, Anime, and Communication

——Where do you get inspiration for your research? You mentioned that you like science fiction. Is that a source of ideas for you?

Science fiction and anime often provide hints when I think about developing and using an idea.[39] I think anime may have made a more significant contribution. I still enjoy watching anime, a habit my wife disapproves of (laughs).

As for inspiration, you hear stories about people being struck by inspiration when eating or taking a leisurely walk. It is similar for me, but I have my own routine. Be it a question or whatever else, when I come up with something that needs to be solved, I have an image of "sinking it." It is like having a swamp in my head and letting the problem sink there. After a while, something that looks like an answer may pop up. If that does not solve the problem, I will push it back into the swamp. In this way, countless things have been forgotten and rotten (laughs), but there are also rare occasions when something I have forgotten comes to mind again. For example, the idea for resource sharing was like that. When I wonder what I should do, I try pushing it back into the swamp in my head. As for when something like an answer comes out of the swamp, in my case, it tends to happen while I'm talking. When I am talking with someone, an idea suddenly comes to me, and I think, "Ah, I can do it this way." Sometimes I think about it alone, but if there is anyone nearby that I can talk to, I explain it all to them without delay, even though it may be an annoying experience for them. They may give me feedback, and in this way, I solidify the idea.

38 https://www.nihu.jp/ja/database/source_map

39 Isaac Asimov's "Psychohistory" (https://en.wikipedia.org/wiki/Psychohistory_(fictional)) provided inspiration for the application of big data to area studies.

——So, talking to people and general communication is also important.

Exactly. The H-GIS Research Group, a group of about 20 humanities and informatics researchers, has met about once every two months for over fifteen years. The meetings have no specific theme. We talk about problems we are having, present new and interesting materials or any new ideas that have emerged, and sometimes, members prepare summaries. However, most of the time, it functions as a space where we just talk a lot. This research group has become a valuable place for me to communicate with others; I often discover ideas that are useful for my research or find ways to advance my stalled research.

Through discussing various ideas in the research group, I have realized that although we may say the same things, humanities researchers and informatics researchers sometimes have different images in mind. For example, once in a mixed researcher group, we tried to plot incident data from a neighboring countries on a map. The locations of the incidents were identifiable to the extent that we could assign latitudes and longitudes. However, without knowing the administrative borders, how could we draw the map? The informatics researchers assumed that since there was no information between the two locations, it was void, and they took the bisector of the points where the two incidents occurred as the border. They thought that if additional information was obtained later, they could modify borders accordingly. The humanities researchers, however, realized that boundaries often follow natural topography and tried to draw the border referencing the mountain ridges and rivers. It is fun to recognize these differences in how we see things, and it can also lead to new research. In fact, the research topic of managing ambiguous borders was adopted as a project under the Grants-in-Aid for Scientific Research.

The H-GIS Research Group is very active in identifying new research themes by polishing ideas that come out of discussions and applying for Grants-in-Aid for Scientific Research. This group also organizes national and international conferences, inviting people from within and outside the university to join our research group.

■Post-Retirement Activities: Continuing Big Data Research and Operating System Development

——What kinds of research activities do you plan to undertake after you retire in March 2023?

Fortunately, the Grant-in-Aid for Scientific Research (A) “Development of Evidence-Based Quantitative Area Studies” will continue until 2026, so I would like to continue my research on big data.

Beyond that, if I really have nothing left to do, I would like to create an OS[40] by the elderly for the elderly. The latest OSs are nice, but they are too complicated to use. I want to make one that is simpler to use. The image I have now is similar to BASIC of the old PCs. It was both a programming language and an OS, and we could do everything from computer control to programming, just by using BASIC. If I could return to that simple world again, it would be a great OS for older people because we would not have to learn many extra commands. I would like to work on it.

(Interview conducted on January 10, 2023 in the Seminar Room on the 2nd floor of Inamori Foundation Memorial Hall, Kyoto University)

40 An operating system, or software that controls the operation of a computer and directs the processing of programs, such as Unix, ↗ Windows, and so on.

Center for Southeast Asian Studies, Kyoto University

An Interview with Professor Shoichiro Hara to Commemorate his Retirement

Pioneering New Fields with Informatics Knowledge and Methods: The Trajectory of Research, Development, and Practice

—Reflecting on a life in research

March 29, 2023

Center for Southeast Asian Studies, Kyoto University
46 Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501 JAPAN
Tel: +81-75-753-7302 Fax: +81-75-753-7350