ISSN 1349-6875

No. 23, October 2007

(This English language edition of NII Today corresponds to No. 37 of the Japanese edition)

# NII Today

National Institute of Informatics News

Special Feature:

## Creating Fusion from Transdisciplinary

The Genome as the Keystone of Integration

Fast, Accurate Determination of Molecular Structure with Efficient Algorithms

The New Value in Information Linkage

Building a Proud Heritage: NII Graduate Education


# *Asao Fujiyama*

Professor, National Institute of Informatics
Professor, Graduate University for Advanced Studies



# *Masago Minami*

Senior Editor, The Yomiuri Shimbun (Tokyo head office)


## NII Interview: Asao Fujiyama + Masago Minami

# The Genome as the Keystone of Integration

## Merging Life Science and Informatics

**Minami:** Taking the integration of life science and informatics as a starting point, I would like to ask you about your area of expertise, genomics. The Human Genome Project, which represented a major step, could not have been accomplished without computers either, could it?

**Fujiyama:** When tangible work on the Human Genome Project started around 1989, a number of databases pertaining to DNA already existed. The DNA Data Bank of Japan (DDBJ), GenBank in the United States, and the European Molecular Biology Laboratory (EMBL) database are typical examples. Considering the volume of human genome data that was anticipated, it was apparent from the outset that computers would be necessary. But the extent of the need for them was beyond all imagination.

**Minami:** Compared with the situation then, the information environment supporting research — computers, communications, and so on — has changed greatly, hasn't it?

**Fujiyama:** In the 1980s large volumes of data were shared by putting information on magnetic tapes and then sending them off by airmail. The acceleration of communications and network formation are a huge plus for users in terms of the advancement of their research.

**Minami:** When did specialists in informatics and specialists in biology begin to cooperate closely?

**Fujiyama:** Back in the early 1980s some biologists were also writing computer programs on their own as the need arose. In the eyes of informatics specialists, though, their efforts seemed unsophisticated.

### CREATING FUSION FROM TRANSDISCIPLINARY

The sequencing of human genome was generally completed at the very opening of the twenty-first century. Life science is now in the age of informational explosion. In particular, the proliferation of data related to DNA and proteins has been astonishing. Databases describing their interaction and relationships are also increasing as are research papers on these subjects. Efforts to control terminology and concepts, which differ depending on the area of specialization, and to systematize knowledge for computer-based storage and calculations are in progress so that people will be able to extract whatever they need from the massive amount of information and knowledge available. These endeavors ought to not only benefit the development of research but also contribute to the dissemination of life science knowledge. The relationship between life science and informatics is such that the two are becoming inseparable.

Gradually this task was left up to the professionals. But even when biologists and information scientists suddenly came face to face, they weren't able to converse. It was all they could do to try to share each other's language and attempt to communicate. There was even debate as to whether the marriage of biology and informatics was feasible.

**Minami:** Is the scenario of a smooth relationship unfolding as envisioned?

**Fujiyama:** Presently a number of universities serve as places that offer training in both fields from the outset. When you perform genome research, you work with information that pertains to the entire spectrum of living things, examine relationships among them, compare the genomes of different organisms, and so on. So you're forced to use computers. The development of people who are capable of doing informatics work in this field has gradually been taking place. Additionally, with a growing need to put the genome and genetic information to practical use at the corporate level, human resources in bioinformatics will be sought not only by universities but also by pharmaceutical makers and other companies. The integrated field of bioinformatics is likely to take hold in this way.

From the very beginning, though, integration is not a thing that is easily accepted. When something is totally new, nobody can gauge its value. As a result it probably won't readily gain general acceptance.

**Minami:** Was that the case with genomics?

**Fujiyama:** "That sort of thing isn't science" was the typical comment. Now, though, genome information is there for use by everyone, and you can't write a paper without using it.

In the case of molecular biology, it probably took about 20 or 30 years to become solidly established. Around the time I was a university student — from the latter part of the 1960s to the first half of the 1970s — early molecular biologists were not accepted by mainstream academia and were struggling hard at schools in outlying areas. At the time molecular biology was a field that integrated chemistry, physics, and biology. The people involved in this field, though, did not particularly perceive it in that way, and they proceeded with a determination to establish a new academic domain. . . . Today the Molecular Biology Society of Japan (MBSJ) has become an organization with a sizeable membership of 8,000 and is Japan's largest academic society in the field of basic biology.

**Minami:** Information also serves as an instrument for promoting integration, doesn't it?

**Fujiyama:** That's right. Genome data have been made available, and anyone can access this information free of charge without restrictions via the Internet. So this information is playing a significant role in the research activities of life science specialists as well as in the work of researchers in other fields.

Genomics researchers consist of people who have a diversity of backgrounds, including medical scientists, botanists, and zoologists. Plus, researchers in a broad range of fields, for instance, engineering, agriculture, and pharmaceutical sciences, are also engaged in activities based on genome information. When fields differ, communication becomes problematic. One reason for this is the use of different nomenclature for the very same things, including even genes and oxygen. Disparities in terminology could also present a barrier to the integration of multiple databases, genome comparison, and so on.

### Jabion: Japanese biotechnology portal site
### A Japanese-language resource for conveying up-to-date biotechnology information in a way that is easy to understand

Jabion 日本語バイオポータルサイト

Consequently, what we are striving to do now is to utilize information theories, including natural language processing and ontology. We are attempting to systematize terminology and concepts and to create a dictionary that people in any field will be able to use. Because an enormous volume of data is being produced, there is also a need to develop technology for the successful extraction of whatever information and knowledge people are seeking.

**Minami:** Doesn't information also have another role in terms of informing the public about this field?

**Fujiyama:** Political forces in favor of sending out information have gained powerful momentum. In reality, however, the people doing research don't have time for that.

We're creating a site called Jabion, a Japanese-language biotechnology portal. One of the reasons for launching it is our belief that scientific articles in newspapers may be somewhat difficult for the average person. There are also instances when limited page space makes it hard to convey background information adequately. Our idea is to take life science themes that are likely to attract attention within society and provide information as plainly as we possibly can.

### A word from the Interviewer:

It's really something that NII has biologists and that biology and informatics are becoming so integrated. While continuing to conduct experiments, Professor Fujiyama — who has been involved in the Human Genome Project from its inception and has always been a pioneer in this field — has initiated a project for the general public to become acquainted with today's biology via the Internet. The number of newspaper articles on basic science is still quite low, with some data indicating that they account for a mere 0.1% of articles in general. I think that a portal site for researchers to transmit information on their own is extremely valuable. I'm excited about Professor Fujiyama's work, and I plan to keep an eye on its progress.

# CREATING FUSION FROM TRANSDISCIPLINARY

## Fast, Accurate Determination of Molecular Structure with Efficient Algorithms

*Hiroko Satoh*
Dr., Associate Professor, Principles of Informatics Research Division

*Takeaki Uno*
Dr., Associate Professor, Principles of Informatics Research Division

**Transdisciplinary research can give rise to results that difficult to reach by means of individual efforts.**
**However, to get academically superior results, it is necessary to understand the fields of the other team members so as to effectively use their specialized knowledge in the collaboration research.**
**Here is an account of genuine integration of chemistry and informatics facilitated by NII.**

There exist much variety of organic compounds in the world, including those are generated by natural resources and are artificially synthesized. In using these compounds for drugs and other materials, knowing the form of their molecules (their molecular structure) are important, so chemists make every effort to determine the molecular structure by means of various techniques. Of these, nuclear magnetic resonance (NMR) spectroscopy is a powerful method. In particular, with carbon-13 NMR ($^{13}$C-NMR), data concerning carbon skeletons can be observed, so it is essential in the structural analyses. One of the important data obtained from NMR is called "chemical shift", which is a significant clue to elucidate chemical structures since it is influenced by the structure around the carbon atom.

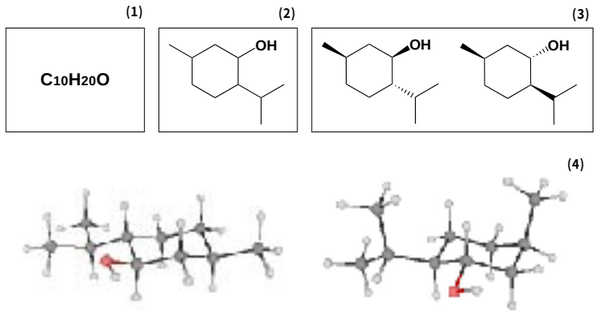



Figure 1: The multiphase representations of menthol (a flavor of mint). 1) Molecular formula. This shows the kind and number of atoms that constitute the molecule. (2) Structural formula (planar). This shows only the atoms and their connectivity (bonds). (3) Structural formula (stereochemistry). With the information of the bond orientation added to (2). The bold line indicates bonds on this side of the plane of the paper while a dotted line indicates those on the other side. (4) Three dimensional structure. (3) and (4) are from left, *l*-menthol and *d*-isomenthol. Their flavors differ slightly.

For many organic compounds, the $^{13}$C-NMR chemical shift data have been obtained. If the data are used correctly and efficiently, they can be a powerful tool for predicting chemical shift and structures. To achieve this, it is necessary to store the structural information with appropriate forms that answer to the purpose of use and to adapt an efficient searching algorithm to it. This is truly a system that requires the integration of chemistry and informatics. Associate Professor, Dr. Hiroko Satoh of NII has developed such a system called CAST/CNMR by collaborating with researchers in various fields.

### Starting to store three dimensional chemical structure data

Generally, chemical structure is represented with coordinates of atoms and their connectivity (chemical bonds). This is useful in visualization, but it is not suited to the tasks of searching and data-processing for advanced use such as prediction. For these purposes, it is necessary to classify the structures according to their characteristics before storing the database. Satoh started this theme at 1996 when she was a Special Researcher in the Basic Science Program, Organic Synthetic Chemistry Laboratory of RIKEN. Her major goal was utilizing a large amount of chemical information effectively to solve problems in chemistry.

While this can be summed up as "molecular structure", there is more to it than that. There are various representations, from molecular formulae showing the kind and number of atoms, to three-dimensional structures (see Figure 1). As actual molecules take three-dimensional forms as shown in Figures 1 (3) and (4), molecules possess the same planar structure may form different three-dimensional structures, making it different molecules. These differences can concern differences of reactivity and property of molecules. Since the same structures can be drawn in different ways, the same structures generally transformed to a unique representation before storing a database. This transforming is called "coding".

Satoh had been taking into account three-dimensional structure as it is essential for the prediction of the chemical reactivity and molecular property. At the same time, Dr. Hiroyuki Koshino, Leader of the Molecular Characterization Team, Advanced Development and Supporting Center of RIKEN (at that time Researcher), had been searching for a database system that could be used in analyses of molecular structures that possess complicated structures as well as stereochemistry, however, no database had been reported that satisfied the requirements for the three-dimensional structures. There are several coding methods for chemical structures, but any one cannot represent three-dimensional structure rigorously. As the two realized that their aims were similar through informal research exchange, they started joint research on the theme of coding three-dimensional structures and predicting chemical shift.

The first stage was developing a coding method for three-dimensional structures, called CAST (CAnonical-representation of STereochemistry). Concerning the development of the CAST method, Satoh looks back, "Development of a flexible coding method that can cover various characteristics both of two-and three- dimensional chemical structures was a tough job". The CAST method codes three-dimensional structures using the dihedral angle (in four atoms, the angle formed by the plane defined by the former three atoms and the plane defined by the latter three atoms). Ring structures are coded as well. Adopting the dihedral angle was a solution through trial-and-error by applying various types of compounds. The benefit point of this approach is that it is a simple way that can be applied to a wide variety of chemical structures.

### Researchers of algorithm joined the team

In early 2000, the second stage, developing a system for 13C NMR chemical shift prediction, called CAST/CNMR was achieved by Satoh developing the software and by Koshino accumulating reliable 13C NMR and structural data from literature as well as his own data. This system uses the CAST coding method to describe chemical structures in its databases. The prediction process involves the steps of coding a partial structure around a carbon atom to be predicted; searching databases with this as the query (search term); and predicting the chemical shift using the hit data.

About the first version of CAST/CNMR, Satoh describes, "In developing its prototype, we focused on developing something that worked, so a problem of processing speed in coding and searching had remained untouched. Also in first version of the CAST method, some problems were involved, for example, it might not give a unique coding in the case of highly symmetric structures." Satoh mentioned these

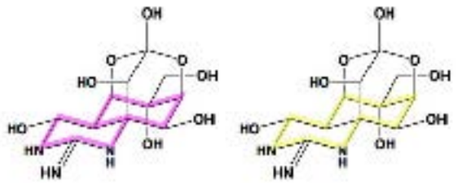


Figure 2: The structure of tetrodotoxin (puffer fish poison). The yellow ring can be formed by combining the pink rings. If rings formed by combinations of several rings are excluded there are eight types, and 12 types of rings if they are not excluded.

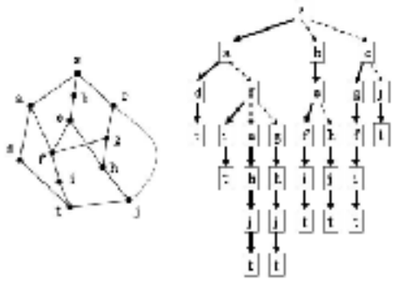
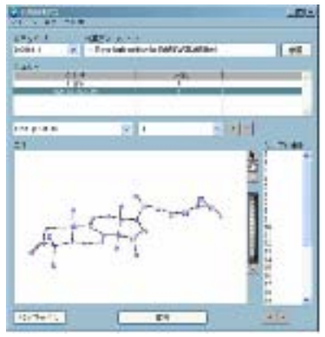
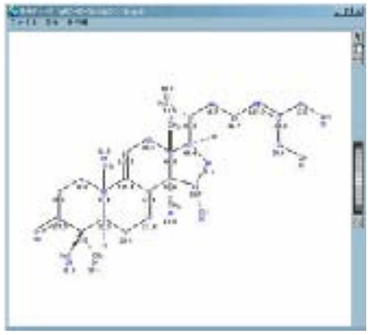


Figure 3: Example of a search using a ring structure enumeration algorithm. Ring structures are found by enumerating the route from s to t in the left figure (the line joining s and t is omitted). The algorithm starts from s, searching by moving to the respective adjacent points. This is shown in the figure on the right. Each of the routes from s at the top to t at the end correspond to each route in the left figure.



The input display.



Predicted $^{13}$C NMR chemical shift values

Figure 5: The graphical user interface of the CAST/CNMR system. If the molecular structure with the condition is specified (left), the 4,000 or so compounds in the database are retrieved, and $^{13}$C NMR chemical shift values are predicted (right). If the user clicks a carbon atom, the molecular data used for the prediction are displayed

problems in her talk in a lunchtime meeting at NII for research exchanging. Associate Professor, Dr. Takeaki Uno recalls his reaction on hearing this. "When I was taking my university entrance exams, chemistry was my second love after mathematics, and I thought I could help with Dr. Satoh's research". Uno's major is developing algorithms. He considers ways of making processing faster by changing the design of the way of the computation than the hardware. Uno called on his algorithm research colleague Dr. Satoru Iwata, Associate Professor of Research Institute for Mathematical Sciences of Kyoto University (at that time Associate Professor of Tokyo University), and the extended team started a joint research from 2003.

Uno and Iwata identified all the problems. When coding planar structures, an atom is chosen first as the starting point, atoms connecting to the starting point are chosen second, atoms connecting to the second atoms are chosen, and so on in order. However, if there exist more than one equivalent atom, a rule is required that strictly defines the order to choose to generate a unique coding.

Solving this kind of problem was what the algorithm researchers were good at. "If more than one order to choose atoms are possible, all possibilities should be enumerated. Since the string of CAST codes can be considered as a string of characters, the problem would be solved by sorting the generated strings in the lexicographical order. With this method, the unique code will always be assigned to the same planar structure, whoever tries it", says Uno. This information scientists' ways of thinking is fundamentally different from the chemists', focusing on types of atoms, bonds, and functional groups. The new algorithm improved the precise of retrieval.

In addition, the processing time for coding was reduced. The first version of the coding method took long time for coding due to repeated calculation, therefore the researchers looked into algorithm theory for methods of speeding it up and discovered logic that did not require recalculation. Through their effort, the processing speed was increased by more than orders of magnitude, for example, it takes within three minutes for coding a molecule consisting 400 atoms.

## The ideal form of integration

As mentioned above, $^{13}$C NMR chemical shift is influenced by the structure around the carbon atom. Ring is one of the important structural attributes that may influence the chemical shift. This is why the CAST method codes the ring structures as well.

However, recognizing ring structures included in a molecule is not so easy task for computer. For example, in order to enumerate the 12 ring structures in the molecule in Figure 2 without omission and without duplication, a mathematically based algorithm is required. In the field of informatics, an algorithm for enumerating ring structures was developed in the 1970s. However, for chemical shift prediction, efficient recognition of only important ring structures is required. Therefore, what Uno worked on was an algorithm for enumerating only rings that are not combinations of other rings.

As Uno looks back, "In 2003, Professor Satoh asked for some advice over a cup of tea. Earlier algorithms found all the rings including ones formed of multiple rings, so processing took a long time. But by adapting just one other logic element, we could improve it relatively easily" (Figure 3). By this improvement, even molecules with complicated ring structures can be calculated in just a moment.

Says Satoh, "Adapting these algorithms to the CAST/CNMR system enhanced its usefulness significantly." By taking into account appropriate ring information, the prediction accuracy for chemical shift improved further (Figure 4). And according to Koshino, "The CAST/CNMR system has been applied to revision of published structures by detecting inconsistency between the reported and predicted chemical shifts."

An evaluation version with a GUI (graphical user interface) has been ready (Figure 5), and Koshino has been conducting validation and evaluation. Satoh says, "The next stage of the CAST/CNMR system will be providing other experts of structure analyses for evaluation to further enhance the quality of the system."

Uno was well satisfied with the joint research; "If algorithm researchers like us carry out research on our own, the problems that we try to solve will be limited to what theorists can come up with. But unlike invented problems, real problems have an interesting twist. It's nice to be presented with a genuine problem like this and to find the solution. It really brings some extra depth to the research."

On the other hand, according to Satoh, "In processing chemical data, we always have to face various and complicated chemical problems. If you don't persevere in resolving each one, you won't get scientifically valuable results. We could get successful results in this joint research, because Dr. Uno and Dr. Iwata, and Mr. Shungo Koichi, a Ph.D. student at Tokyo University, who developed the program, recognized this chemical complexity and showed great tenacity." This ideal integration of informatics and chemistry is set to continue producing useful research outcomes.

(Written by Seiko Aoyama)

The result by using data having the same planar and three-dimensional partial structure as the blue partial structure.

The result by using data having not only the same planar and three-dimensional partial structure as the blue partial structure but also the same ring structure as the yellow rings.

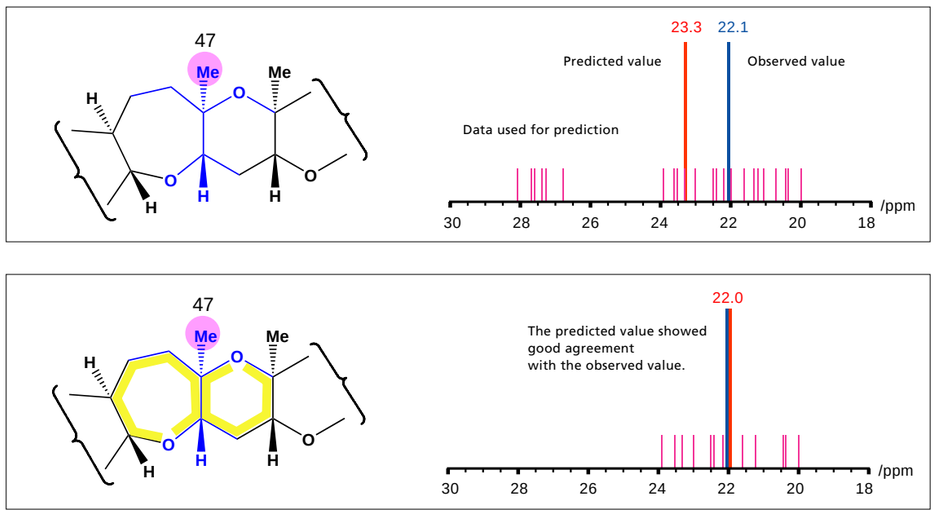
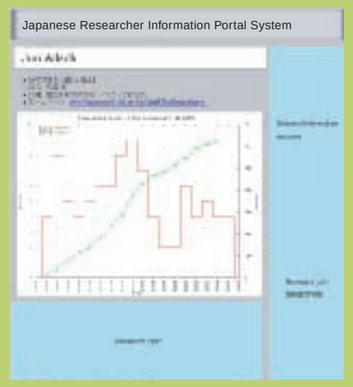



Figure 4: Example of improved NMR chemical shift prediction using CAST code including ring structures. The chemical shift of the 47th carbon atom, which is included in the methyl group shown as pink circle, in brevetoxin-B, a marine natural product.

# The New Value in Information Linkage

**Today's researchers are looking for new value in information by using "information linkage"—assembling information about the same item or person and placing it under integrated control. Drawn by the theme of information linkage, what kind of new value are they aiming for?**




Google Maps

An example of a "Japanese Researcher Information Portal System" constructed based on information assembled from a variety of sources. The information can be displayed in lists, graphs, and maps. The graph at left shows the relationship between numbers of participating projects (red bar) and the accumulated list of publications (green broken-line). The map at right shows the distribution of top 100 joint researchers. Director of NII's Cyber Science Infrastructure Development Department, Professor Jun Adachi is a member of the Transdisciplinary Research Integration Project.

Akiko Aizawa
Professor, Digital Content and Media Sciences Research Division

"Information linkage" is conventionally known as "account consolidation."(*1) "While it's not very noticeable, it's an extremely important operation," says Akiko Aizawa, leader of the "Collection and Analysis of Data for Large-Scale Linkage" group at the National Institute for Informatics Transdisciplinary Research Project.(*2) Professor Aizawa's original motive for making use of information linkage was as a form of "database-cleaning," namely, "to eliminate duplicate records in databases of academic papers and online catalogs."

For example, when the same family and personal name appear multiple times as an author record, it's important for the user of the database to know whether it is merely a duplicate record, or a different person with the same name. Using information linkage to improve the quality of the database has great benefits for actual use of the data (see accompanying graph).

With the spread of computers and gigantically growing networks, the volume of data not only continues to expand, but it is also uncontrollably disordered and unwieldy. There are indeed innumerable occasions when information linkage is required to organize and order information in response to specific needs.

## Information Linkage and Databases

Not all information linkage is the same, however, and the linkage involved can occur in several different patterns. Professor Aizawa states that "the problem setting can be divided into four levels based on degree of difficulty."

First, as represented by the example posed earlier, one can consider operations performed within a single database. A database is a collection of data assembled in categories and formats for a specific purpose, with the result that it is relatively easy to compare data items and determine whether they're the same (deduplication).

Next is the case of integration across two or more databases. Information linkage ("account consolidation") had unintentionally become known recently after the Japanese pension-records mishandling reports, in which comparison and identification was performed on personal records scattered across databases in different locations, and those which could be determined to belong to a single individual were then integrated. In that way, integration can be performed across the entire database.

Up to this point we've been talking about databases with established data formats. But there are great quantities of useful information outside of databases proper, on the Internet, for example. The remaining problem is how to draw that totally heterogeneous, disordered information into linkage. Namely, the third level is the linkage of information in databases with that outside databases, while operations entirely outside of databases represent the fourth level.

## Toward Concrete Services

Even taken alone, this degree of information linkage has a great deal of real value, but the purpose of the project is to use linkage to assemble information and then to effectively utilize the "something extra" created from that assembled data. Associate Professor Kei Kurakawa, working at the Research and Development Center for Scientific Information Resources (director, Hideaki Takeda) is currently engaged in joint research aimed precisely at giving concrete shape to that "something."

"With the purpose of developing academic and scientific contents services, we are first doing name identification for Japanese researchers' data. This will be of benefit, for example, when looking for a joint researcher, or when evaluating researchers' performance." Name identification is being performed on the Ministry of Education, Culture, Sports, Science and Technology's researcher number database for grants-in-aid for scientific research, together with databases of researchers at individual universities; this can thus be called a concrete implementation of the "second level" of linkage. Masao Takaku, a researcher at the Transdisciplinary Research Integration Center, is focusing on the same research-grant database to reformat the data, focusing on links between individual researchers. By paying attention to joint researchers, and the relationships between team leaders and their associates, Takaku says "We may be able to understand, for example, things like who should be made team leader in order to facilitate receiving research grants." The kind of useful data produced depends on the kind of information linkage employed.

Foreign student interns are also active in Aizawa's research group. Dang Bac Van is a master's student at Vietnam National University - Ho Chi Minh City, identification results, where he is engaged in research on the extraction of information from text data. This is an attractive research theme with broad practical applications. For example, by extracting the names and relationships of methods and tools described in an academic research paper, it may become possible to apply them to other fields. Van is currently giving his efforts to the theme of how to extract text data without large investments of human labor.

## Associates Drawn Together by Information Linkage

The various members of Professor Aizawa's research group were not all originally involved in information linkage research, and there were few commonalities between them. Professor Aizawa began from communications, going on to machine learning, optimization, text and natural language processing. Associate Professor Kurakawa, by contrast, studied engineering design science and software design. The only point in common with Professor Aizawa was the fact that he was utilizing the same data. Most of the researchers assembled in the group are like this.

Also, the everyday research landscape is different from the usual image of the physical sciences research environment, where associates are pushed together in the same room each day. Instead, rooms are provided where all members can gather independently to discuss issues as required. Since Van is a student, he has regular meetings with Professor Aizawa, who, he says, "has given me numerous good ideas, thus allowing me to engage in productive research," but in principle, such meetings are held only when considered necessary. But when they do gather, the group members display strong unity. According to Aizawa, "Everyone has a different background, so we are stimulated anew each time we have a discussion, making it a very enjoyable environment for research." Truly, this is a research group brought together through "linkage."

Masao Takaku
Researcher, Transdisciplinary Research Integration Center

## The Goal: Easier-to-User, More Valuable Data

Professor Aizawa states, "At present, we have created a system that can handle information linkage without problem up through the second level. But since human resources are required to verify the identification results,, we run up against the problem of what degree of quality we can achieve at what cost." Many issues are amenable to solution by sharing databases, such as the problem of concentrating research grants on specific individuals, and the problem of pensions noted earlier. The rest is basically up to implementation policies.

When it comes to future directions, Professor Aizawa suggests that "by using information linkage, we have basically come to the point of being able to understand to some extent things we were unable to grasp at all before. We next intend to assess the value of that partial knowledge and create a system that will provide truly useful information to society." On the other hand, she is also considering ideas for new kinds of services. It is thrilling to contemplate what kind of new value will be created by these researchers, assembled under the umbrella of information linkage.

(Written by Tomoaki Yoshito)

Dang Bac Van
Student Intern, International Internship Program

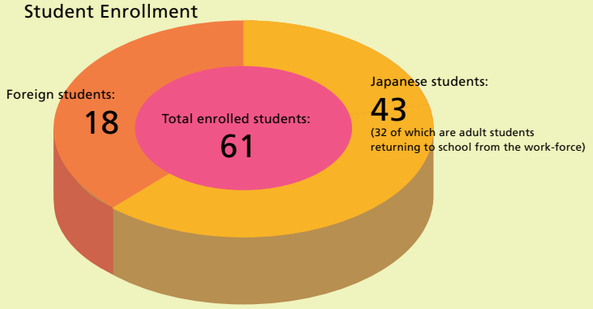
*1. While the same as "account consolidation" in the sense of gathering data together in one place, "information linkage" emphasizes the connections between the data thus gathered.

*2. The Transdisciplinary Research Project called "Creation of Information Space and Information Foundation for Transdisciplinary Research Integration" has three subprojects, and Aizawa's group is included in the first, "Construction of Methods and Knowledge Base for Large-Scale Heterogeneous Information Collection, Analysis, Linkage, and Classification." At present the project is in the third of a five-year plan.

Kei Kurakawa
Associate Professor by Special Appointment, Research and Development Center for Scientific Information Resources

# Building a Proud Heritage: NII Graduate Education

The National Institute of Informatics (NII) has three faces: a center of comprehensive research on informatics, an interuniversity research institute, and an educational institute where individuals receive high-level training in informatics. In order to get a better idea of the current status of the Institute's program of graduate education, I interviewed Professor Noboru Sonehara, acting manager of student recruitment, together with several recent graduates and current students.




NII's graduate school represents one center in the Graduate University for Advanced Studies formed jointly by eighteen national research institutes under the aegis of the Ministry of Education, Culture, Sports, Science and Technology. Some sixty students of informatics currently study within NII's world-class environment.

**Sonehara:** I've asked you all here today to talk about NII's graduate school. I'd like to start by asking for self-introductions, and the reasons you decided to apply here.

**Ohmukai:** I'm studying how blogs and online social networks change our ways of communication. I was in Kyoto through my M.A., but I felt there wasn't enough information for my research there, so decided to come to Tokyo.

**Kajiyama:** I feel it's becoming increasingly important to know how to search for what you need out of the exploding mass of information available. My goal is a search interface that makes the act of searching itself enjoyable. I felt that NII, with its leading status in information retrieval, was the optimum research environment.

**Kanokwan:** I came to Tokyo to study the Japanese language. While searching on the Internet for graduate schools matched with my interest on electronic commerce, I found NII. Some recommendations from NII's students also influenced me to pursue a Ph.D. here. Currently, I have been researching on the factors affecting the success of electronic commerce.

## Ph.D. Course for Study and Research

**Kanokwan:** The research facilities here are modern, and instructors are all professionals. The atmosphere makes it easy to discuss issues with instructors and therefore I never have any uncertainty with regard to my research.

**Kajiyama:** While being in the Ph.D. course means you're a student, at many universities you have responsibility for teaching the younger students and don't have time to devote yourself to your own research.On the contrary, NII has allowed me to fully use my time as a student for effective learning.

*Noboru Sonehara*
Professor, Information Infrastructure, Information Commerce System, Information and Society Research Division; currently researching solutions in digital commerce and the creation of a foundation for information reliability.

**Ohmukai:** In addition, I like the atmosphere of being able to do free research. It's so free, in fact, that some people might feel neglected, but it's ideal for students who want to set their own theme and design their own research program.

*Tomoko Kajiyama*
Completed coursework for Ph.D. in 2007; currently Research Associate in the Department of Human Informatics and Cognitive Sciences of Waseda University's School of Human Sciences.

**Sonehara:** Fundamentally, research is fired by the possibility of free creativity. We're proud of the fact that it is our status as a small graduate school that allows us to offer an open environment for research. In order to encourage our students to push forward with uninhibited research, NII's network and online academic contents are used to publish their research results throughout the world.

**Kajiyama:** NII has the will to publish research results to larger society; they even made a patent application based on my research.

**Ohmukai:** While I was in school, I started a company to develop an RSS (Rich Site Summary) reader, a tool for reading online blogs.

**Sonehara:** Of the eighteen central research institutions within the Graduate University for Advanced Studies, we are the only comprehensive educational institution dedicated to research on informatics. Part of the mission of the faculty and staff here is to disseminate Japan's intellectual and information services strategy and information communications technology strategy, so we are constantly aware of the situation in the larger world outside.

## The Synergy of Diverse Faculty and Students

**Ohmukai:** Another characteristic of NII is the diversity of the people here. Of the students in attendance, one-third are foreign students, and one-half already have post-graduate employment experience. This is a balance unheard of at other graduate schools. Many of the instructors have also experienced employment in the private sector, outside of academia.

**Kajiyama:** That's why it wasn't as closed an environment as an ordinary university, and it represented a better situation for me before going out into the world. My desk was near Ms. Kanokwan's, so we became friends, and I think that we provided each other with good intellectual stimulus.

**Ohmukai:** What I find interesting now is the fact that NII students are all people who accepted the challenge of "changing their environment." This goes not only for foreign students and older students with employment experience, of course, but I think all those who choose the Ph.D. program here have challenging personalities. If you change your location—your environment—your way of thinking also changes. And that is linked as well to changes in your research topics. NII is a gathering of that kind of people, so they're all serious thinkers.

## NII's Broadening Web of Influence

**Ohmukai:** I've been a research staff here for three years, and I feel the environment for research is really tops. I'd like to stay here forever, but on the other hand, to do good research, you also have to maintain connections with the outside world. For the present I'm trying to keep up my contacts while focusing myself here at NII.

**Sonehara:** We tell our graduates we want them to aggressively go out and get jobs in society. The times demand that kind of "centrifugal force." And when I say "centrifugal force," I mean to transmit the things students learn and acquire here at NII to the outside world at large. After all, that should be the mission of Japan's highest institution of learning . . .

*Ikki Ohmukai*
Completed his Ph.D. coursework in 2005 as a member of the first class to complete their program of study at NII. Currently, working as Assistant Professor in the Digital Content and Media Sciences Division of National Institute of Informatics

**Kajiyama:** I've been employed at Waseda University since this April. I'm a woman, and when I consider that the present time is a priceless period that I can use for my own purposes, I first thought I'd like to concentrate on my research. But now I really feel a significance in passing on what I've learned to students.

*Kanokwan Atchariyachanvanich*
A student from Thailand, Kanokwan is a 2001 MSc graduate in Information Management from Asian Institute of Technology (AIT) and currently a third-year student in the doctoral program at this time.

**Ohmukai:** Most of the graduates from NII take employment at organizations where they previously performed joint research as students. That's great because the work they did was clearly recognized, so it's something to be proud of.

**Kanokwan:** I originally thought I would return to Thailand after graduation, but now I would like to continue my research at either NII or a Japanese corporation – any place giving me an opportunity.

**Kajiyama:** Although this has the perfect environment for research, it's unfortunate that the graduate school itself is not very well known.

**Ohmukai:** I like new things, and I think their name value can be made by yourself. I really feel that the things I've done are helping to write the history of NII.

**Sonehara:** With the current era of lowering birthrates, every university is fighting to acquire its share of the student population. We're also trying to create an environment that is amenable for learning, both for exchange students and for older students who are returning to school after years of employment. But enhancing the brand identity and value of an NII education depends on the activity and results produced in society by people like those assembled here today. NII will continue to evolve and improve so long as the two wheels of research and graduate education continue to turn smoothly.

Thanks to all for taking time from your busy schedules to talk with us here today.

(Written by Akiko Ikeda)

NII ESSAY

# What is name identification?

*Akiko Aizawa*
Professor, Digital Content and Media Sciences Research Division, the National Institute of Informatics

All sorts of events occur during the course of peoples' lives—from ordinary things such as births, deaths, marriage and divorce, to adoption and immigration. In all of these cases, various forms of notification have to be made, often dispersed across a wide geographical range. And since these documents are handled by people and therefore liable to human error, misspellings and data input errors can occur.

## The difficulty of linking records

Linking each and every one of these geographically scattered individuals' records is somewhat more difficult than it appears, but it is paramount to civil life because it is the basic data for proving a person's identity and providing the proof by which their family allowance and pensions are paid.

Linking records is the process of checking two pieces of data and deciding whether or not the content conforms. Though this would at first appear to be a task that could be performed easily with computers, this is not actually the case. The Von Neumann-type calculating machine was first conceived in 1946, and whilst the computerization of society subsequently progressed rapidly the essential difficulty of the issue of linking records remains unchanged.

Let's take a look at a good example of this difficulty by trying to find out about a certain person using search software. First, we enter the surname - resulting in a mountain of data about the same name and the names of places etc. Entering additional keywords about the person's affiliations etc. should help to narrow down the scope of the search. But in many cases, establishing whether or not the content then displayed pertains to a certain person is a task that eventually has to be done through manual verification. In other words, the role of combing records is also a question of verifying descriptions - whether or not the person in one record is the same as the person in another record.

## Verification of statements, and costs

When the role of linking records boils down to the verification of people's statements, one cannot help but be reminded of the "name identification " phrase that filled newspaper columns and became a contemporary buzzword during the recent public pensions problem in Japan, in which millions of payment records were lost.

"I haven't been paid the pension I was supposed to be enrolled in," "Who do the disappearing pensions belong to?" These are simply a question of verifying the veracity of statements about people, and 'identifying their names' to see if they are the same people.

However, this difficulty, which has remained unsolved for over half a century, appears to posses all the essential perplexities of data processing. Vagueness in notation cannot be avoided as long as human intervention is involved, and we have to approach data processing with the premise that this vagueness exists. In doing so, what is important is information about who carried out the verification and in what manner: in other words, the clarification of where the responsibility lies. It is with respect to this responsibility that the considerable costs of verification are entailed.

However, these costs should be considered an investment in the additional merit of clarifying verification and the locus of responsibility, rather than merely as compensating for the losses caused by incomplete information. Therefore, treating unverified statements as if they have actually been verified is liable to incur chaos, which is exactly what we have witnessed in the case of the pensions fiasco.

Furthermore, the fantasy of dramatic cost cuts through automated data processing is, as half a century's research shows, little more than an overestimation of the value of computerization.

The issue of linking records is an excellent example illustrating that the knowledge we automatically obtain from computers and the knowledge verified by experts are both assets, and both essential.

Weaving Information into Knowledge
NII
**NII Today No. 23, October 2007** (This English language edition of NII Today corresponds to No. 37 of the Japanese edition)
Published by: National Institute of Informatics, Research Organization of Information and Systems
Address: National Center of Sciences 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
Chief editor: Yoh'ichi Tohkura Cover illustration: Makoto Komori Photography: Shuichi Yuri Design: Kotaro Suzuki Production: Sci-Tech Communications Inc.
Contact: Publicity and Dissemination Team, Planning and Promotion Strategy Department TEL:+81-3-4212-2135 FAX:+81-3-4212-2150 e-mail: kouhou@nii.ac.jp