ISSN 1349-6875

NII Today

National Institute of Informatics News

No. 22, July 2007

(This English language edition of NII Today corresponds to No. 36 of the Japanese edition)

Special Feature:

Beyond the "Information Explosion"

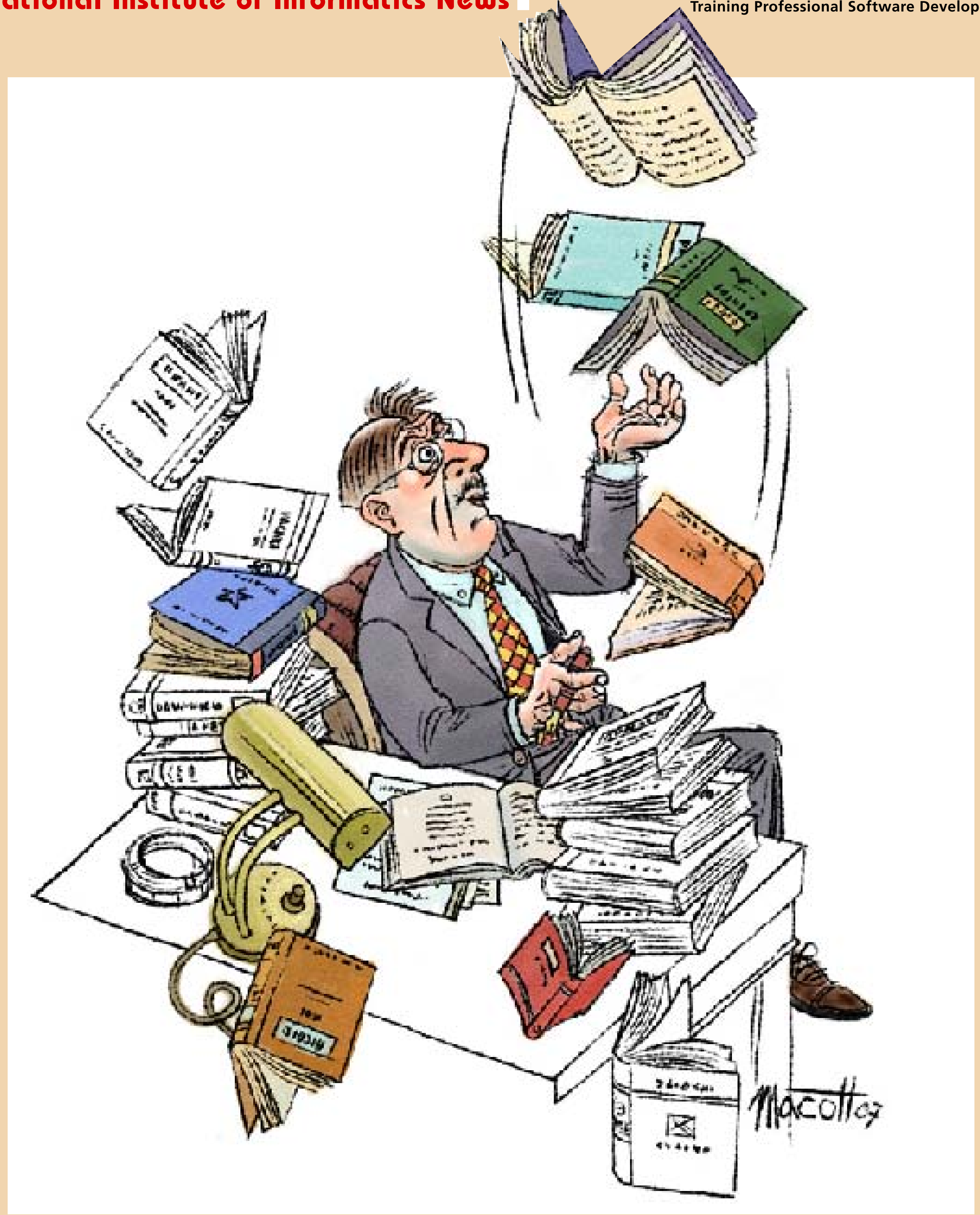
From Search to Association: Information Technology that Brings Inspiration

New Information Retrieval Technologies Using Videos

Making Cultural Heritage Objects Useful through Subjective Annotations

Academic Portals for the "Information Explosion" Age

Training Professional Software Developers



Akihiko Takano

Professor and Director, R&D Center for Informatics of Association
National Institute of Informatics (NII)

Mariko Takahashi

Science News Editor, The Asahi Shimbun

NII Interview: Akihiko Takano＋Mariko Takahashi

From Search to Association

Information Technology that Brings Inspiration

Beyond the Information Explosion

Takahashi: I saw that the May 11, 2000 issue of the journal *Nature* featured an article entitled "Souped-up Search Engines" in which associative search (*1) technologies were discussed along with Google.

Takano: At that time, I was working at Hitachi, Ltd., and the "DualNAVI" associative search interface we developed was mentioned in the article. The essence of DualNAVI was extracted as independent software known as GETA or Generic Engine for Transposable Association. GETA was released as open source software, and it can be used freely by anyone at no charge.

Takahashi: Do you think that, given the opportunity, associative search could have been rival to Google?

Takano: Even now, I still believe that association is a technology that comes after Google.

Google's truly exceptional quality is that, on its own, it has been able to create mechanisms and equipment that can gather enormous amounts of information on the web and create page rankings accurately without confusion. Another exceptional aspect of Google is its effectiveness in attracting the necessary financing and collecting web data in an unimaginable scale.

On the other hand, we have had a simple belief that what we have created is superior to Google in some sense, and we still feel that way. With an explosive growth in the amount of available information, the relevant information to any given request will become impossible to be returned as a whole . The question of how to rank the search results and so on comes second to the problem of how to get a broad overview of the search result. In associative search, the related keywords can be extracted from the data gathered, and they help us to get a broad overview of the data content.

For this reason, it is only now, as we enter what is being called the age of "information explosion," that associative search is demonstrating its true value. In other words, I am convinced that associative search as a technology was developed from a perspective that foresaw the walls that Google is now hitting. Now that the core technology is complete, our focus will be on maintaining the same perseverance as Google in determining how to use this technology to provide effective and interesting services.

Takahashi: One of the interesting things about this technology is that the basic research was done at an industrial environment and the applied development is being pursued at a national institute. What made you decide originally to do research in associative retrieval?

Takano: It started back in February 1996 when I was at Hitachi Advanced Research Laboratory. Several small groups were reorganized into a single research team, and I was chosen to lead that team. The team included one person who was studying the measures to evaluate the proximity among documents in terms of content (information retrieval), and another person who was studying word sense disambiguation with regard to words like "bank" through contextual analysis (natural language processing).

Up to that time, I had had no interest in anything other than writing beautiful programs in an elegant programming language. To me, their approaches seemed like two sides of the same coin. I felt that evaluating the proximity of documents from the frequency of the words on the one hand, and determining the meaning of words from the surrounding context on the other, were inextricably linked. The coming together of three of us was the catalyst that inspired the concepts of associative search and computation for association. Then an outstanding software engineer joined and he turned this concept into reality.

Takahashi: How did these technologies developed at Hitachi become open source?

Takano: At the end of the 90s, the boom of basic research at industries subsided, and the pendulum began to swing back to the more practical research. We were moved to Central Research Laboratory, and it was at that point that we had to create an environment where researchers in our team could continue research in each individual areas even if they went to different workplaces — for example, even if they quit Hitachi. Otherwise their spirits and morale would be shaken and ultimately no achievements would be possible.

Of course, Hitachi had put a lot of resources into the research on association, and it deserves to own DualNAVI, the killer GUI for association we created. After we made it clear by applying several patents on it, we persuaded our boss that making the associative calculation engine an open source would expand the field itself and ultimately benefit Hitachi. In February 1999, just before we moved to the Central Research Laboratory, we submitted a proposal to the Information-technology Promotion Agency (IPA) for an Innovative Information Technology Incubation Project, and the proposal was approved. The outcome of this research project was the aforementioned GETA. I believe this is the best decision I made as a manager from my company days (laugh).

IMAGINE

"I believe association is the technology that comes next after Google."

In January 2001, I joined NII, which maintains the union catalogue of university libraries nationwide. The database covers 1000 libraries and includes eight million records of books and journals. I thought this database is ideal for associative search to experiment with, and created Webcat Plus(*2). Now, Google finds Webcat Plus as the top-ranking result for the search by "association". And another service Shinsho Map (*3) is the top result for "shinsho (paperback)".

Takahashi: Meaning that these technologies now become widely used by people. What is the next step for associative search?

Takano: The current tidemark of associative search is IMAGINE (*4). It facilitates users to combine several reliable information sources, and interact with them seamlessly by associative search. A wide perspective from various sources realizes more balanced way for navigating within digital space. We believe it has the potential to replace keyword searches in the future. We are currently busy to develop a multilingual version of IMAGINE that can help association go beyond linguistic barriers. As our association does not rely on translation, it goes beyond culture. We hope it would make it possible to determine the correspondence between different cultures such as China, United States and Japan.

The competition among search engines to get the greatest number of hits will soon be over. And we will enter an era of association technologies, with which users get inspiration based on reliable information.

✎ A Word from the Interviewer:

Associative search is truly a fascinating technology. You lose all sense of time as you follow related documents and keywords one after another. It's like an intellectual amusement park. My only concern is that, if you spend too much time in this amusement park, you'll have no time to read the document you were originally searching for. There's a danger that you'll always think: "There must be a better reference" and you'll never settle on anything. The goal should be to provide a service that answers the needs of the ordinary person. The CEO of Google told The Asahi Shimbun: "Our ultimate goal is to provide a service that enables people to truly understand the meaning of the information they search". I hope NII won't allow itself to be outdone in this regard.

(*1) What is associative search?
If you provide a free text as a query, the system returns a collection of similar documents with ranking and a list of "related words" which summarize the content of them. These related words are helpful to stimulate the users to get new ideas or concepts. Users can select some documents and related words and use them as a query for the subsequent searches. This facilitates users to start with just a vague idea and gradually approach to the desired information.

(*2) Webcat Plus — Information service with associative search
Search service for the union catalogue of university libraries nationwide which is run by NII.

(*3) Shinsho Map
About 10,000 paperbacks are classified into 1,000 live topics. Users can search related topics and related words by associative search. Users can enjoy the theme surfing among related topics.

(*4) IMAGINE
This service provides the associative search for multiple databases simultaneously, including Webcat Plus, Shinsho Map, Book Town Jimbou (listing 310,000 books from 160 used book stores in Tokyo's Jimbocho area), Cultural Heritage Online (listing 5,000 items stored at museums and art museums), Wikipedia and so on.

Academic Portals for the “Information Explosion” Age

National Institute of Informatics (NII), in cooperation with universities and other entities, collects a comprehensive array of academic information created at research institutes. By reorganizing this information in an easy-to-use form, NII is working to develop an academic portal site to meet the diverse needs of the “information explosion” age.

In order for Japanese academic research to become more competitive, the academic content needed by the academic community must be secured, and it must be possible to add value to this content and disseminate it to the rest of the world. However, the number of academic papers being submitted has increased dramatically, and the price of overseas academic journals has also skyrocketed. Moreover, it is not only academic papers that are needed as academic content. According to Koichi Ojiro of the Cyber Science Infrastructure Development Department at the National Institute of Informatics (NII), important content also includes "experimental, statistical and observational data that serve as the basis for educational materials, academic papers and so on that are prepared and publicized by researchers, as well as software tools developed as research aids, cultural property archives and so on." This type of content is increasing at an even more explosive rate than academic papers, but very little of it has been recorded and stored in a form that enables it to be searched and used.

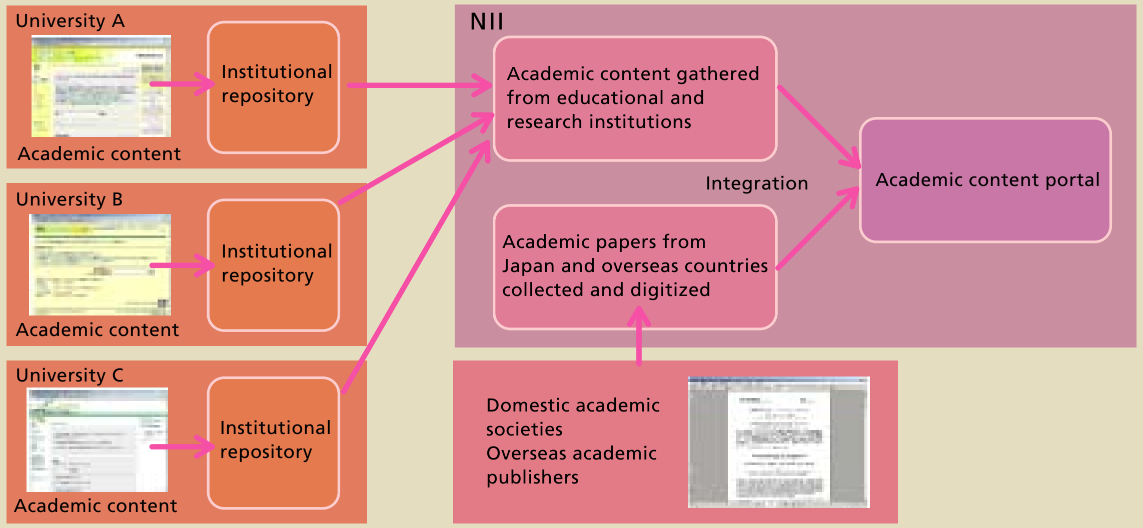
NII Helps Construct Institutional Repositories

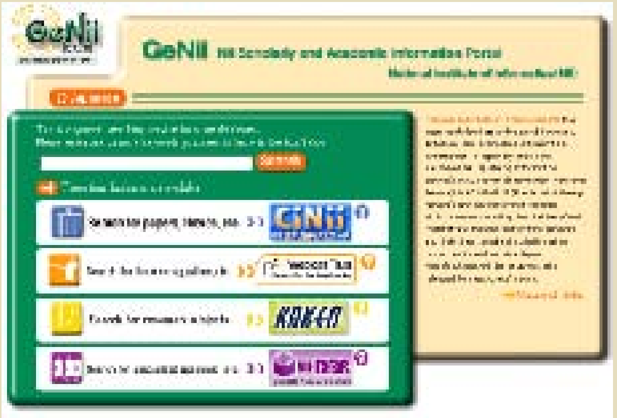
For this reason, NII has been helping to construct "institutional repositories" — digital libraries in which universities and the like can manage in an integrated manner the academic content produced through research at such institutions. The institutional repositories created in this manner can be used independently or used in conjunction with other institutional repositories via the Internet. For this reason, the bibliographical information (metadata) and so on must be prepared in accordance with the same standards.(*1)

Beginning with the "Project for Experimental Implementation of Software Building Institutional Repository (NII-IRP)" implemented in fiscal 2004, NII has provided assistance for the construction of institutional repositories. NII has also provided technical information for the software needed for construction and training of maintenance staff. As of April 2007, consignment contracts had been concluded with 57 universities, and completed institutional repositories had already been made available to the outside world at 41 of these universities. Although some universities in Japan have constructed institutional repositories without the assistance of NII, Mr. Ojiro says that "as long as the same standards are adhered to, people will be able to cross-search the systems without being aware of any differences."

In addition, NII has been cooperating with academic societies in Japan to construct a digital archive of academic papers. Moreover, in cooperation with consortia of university libraries, NII also collects data from overseas academic journals. Recently, joint subscription agreements for e-journals have been concluded with the German publisher Springer and Oxford University Press, and these e-journals have been made available together with the domestic Japanese digital academic paper archive. (As of April 2007, 6.3 million works were available.) In addition, NII has also provided assistance for the Cultural Heritage Online project. This project, conducted by the Agency for Cultural Affairs and the Ministry of Internal Affairs and Communications, is designed to make available in digital form tangible and intangible cultural assets, including museum collections and the like.

Collection and integration of academic content


University A
Institutional repository
Academic content
University B
Institutional repository
Academic content
University C
Institutional repository
Academic content
NII
Academic content gathered from educational and research institutions
Integration
Academic papers from Japan and overseas countries collected and digitized
Academic content portal
Domestic academic societies
Overseas academic publishers



Top page of GeNii, the NII academic content portal. In addition to the four databases shown here, NII also offers academic paper information that integrates digital journals from overseas publishers and makes cultural heritages and the like available online.

Devising Methods to Provide Content

The content collected in this manner takes many forms, including text and PDF files, HTML, graphic files and so on. For this reason, a method is needed to organize and integrate these many different formats. For example, in many cases the same paper has been derived from multiple sources, and deduplication is needed. Technologies developed by NII researchers are helping to streamline this enormous quantity of pre-processing work.

In addition, the information must also be provided in a form that will be easy for people to use. "Even if we provide a mechanism for simple cross-searching at the text level, this would not attain the ideal level desired by researchers," says Mr. Ojiro. "The problem we need to resolve from this point on is how, and in what form, to extract the necessary information from content that is heterogeneous in many ways and then turn it into 'knowledge'. The new search technologies and information presentation techniques that have been developed by NII researchers will be effectively incorporated to make academic data 'visible.'"

Some of the services already being provided by NII have incorporated the results of certain research projects on an ad-hoc basis. In the future, an effort will be made to make services more advanced and more user-friendly, primarily at the Research and Development Center for Scientific Information Resources (http://www.nii.ac.jp/cscenter/) set up in 2006 within NII. Both researchers and operational department staff from NII participate in activities at the Center. Plans call for the achievements of NII researchers to first be made publicly available as test services. After needs and so on have been studied, these achievements will be incorporated one after another into actual services.

Another issue is coordination among the existing services that have been provided up to now. "For example, in current services that provide the achievements of grant-in-aid research, it is possible to search for research reports and display a list of the titles of papers that describe research achievements," says Mr. Ojiro. "But at present that is all that can be done. In the future, we want to create a mechanism that will enable someone to click on the title of the paper and download the paper from the NII digital research paper archives."

Seeking to Provide Value that is Distinct from Google

In the online world, NII services for collecting and providing academic content are subject to the same competition as the services of private sector companies. The one that engenders the most concern is the American firm Google. Google Scholar can search for academic papers, and Google is increasing its service options with links to the U. S. Patent Office and various other government agencies. And the company has been improving its search functions and user interface with surprisingly rapidity.

And, truth be told, NII also linked up with Google in April 2007 with the aim of expanding its user base, to enable Google users to access the NII database both from Google Scholar and via Google web searches. According to a recent analysis, by the end of 2007 access from Google may exceed that from the NII top menu. To the academic societies and educational and research institutions that provide data, an increase in the number of hits is undoubtedly good news regardless of the route users have taken to get there. However, NII cannot be satisfied with this state of affairs. Mr. Ojiro expressed the desire to develop a service with a difference, "one that does not merely rank the wheat and chaff of content based on a single measure but provides high-quality content in an appropriate form that matches the advanced needs of the academic community."

(Written by Atsushi Saito)

(*1) The standard for interoperability among institutional repository systems

The Dublin Core Metadata Element Set (DCMES) is the standard for the bibliographical information (metadata) used for describing digital resources. In addition, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been prepared as a protocol for automatically gathering and sharing metadata from multiple digital archives. If these two standards are followed, metadata can easily be exchanged between institutional repositories in various locations that are connected to the Internet.

Training Professional Software Developers

The Top SE Project is a joint industry-academia project whose goal is to train technology leaders in software development. The project's third year ended this spring with its first graduating class of 12 students completing the program and being sent out to the front lines of software development.

Most university graduates who majored in information science feel that some of their coursework does not prove to be very useful. This was one of the findings contained in the "Report of the Field Survey on IT Education at Universities" issued by the Ministry of Economy, Trade and Industry in March 2004. The report concluded that, for example, the course on Formal Language and Automata "cannot be said to be unnecessary but should be studied to determine if it should be made an elective course."

Fusing Information "Science" and "Practice"

Professor Shinichi Honiden, project manager of the "Education for Intellectual Product Development through Science", also known as the Top SE Project (*1), describes this situation as "unacceptable." Professor Honiden also teaches at the National Institute of Informatics. "There is no waste in university courses," he emphasizes. "Partly due to the fact that deadlines are tight, there's no "science" in the software development in industry; it's a crash program. At present, there is a lack of people who have a thorough knowledge of information science and can do high quality design work."

At the same time, although universities have cutting-edge software tools and techniques, they do not skillfully put these to practical use.

In other words, "there is no science in industry, and there's no hands-on training at universities." The goal of the Top SE Project is to fill in this gap by training "superarchitects" who can do high-quality software design for new and very difficult development problems.

Professor Honiden gives a press conference following the course completion ceremony.

Training People Who Can Create "Blueprints"

How exactly can such "superarchitects" be trained? To answer that question, it is first necessary to determine the skills they require. Professor Honiden says these skills can be summed up in one word: modeling. "In the course of solving problems, "superarchitects" must be able to determine the essentials and figure out what tools to use and how to use them. In other words, they must have the ability to create 'blueprints.' This skill is important at each stage in the software development process: planning, development, operation and maintenance."

For this reason, the Top SE project stresses training in software tools. Once students have a knowledge of information science, the emphasis should be on how to use tools skillfully. This does not mean becoming proficient in the use of only a single tool. In the field of information science, in which new tools are created one after another, choosing the most appropriate tool for solving a particular problem is also a necessary skill.

In the Top SE Project, training is conducted based on a joint industry-academia organization. NII and universities provide knowledge of information science and state-of-the-art tools, while the IT industries provide practical case studies. Instructors are dispatched not only from academia but from industry as well. When the project was inaugurated, there were five cooperating companies. As of April 2007, this number had grown to 14, and during this fiscal year the number is expected to increase still further to some 20 companies.

Curriculum

Process \ Issues	Security —Development of high-quality software with no bugs—	Reliability / efficiency	Changeability —Maintenance—
Requirements Analysis	Requirements Analysis Security Requirements Analysis		
Design	Formal Specifications (Security)	Formal Specifications (Foundations) Formal Specifications (Applications) Verification of Design Models (Foundations) Verification of Design Models (Applications) Modeling and Verification of Concurrent Systems Verification of Performance Models	Agile Development Aspect-Orientated Development Software Patterns Component-Based Development
Implementation	Verification of Implementation Models		
	Testing		
	Program Analysis		

Training of professionals is pursued in accordance with a well-designed program.

Personnel Training and Development of Teaching Materials

Top SE courses are held once each week for two periods. One credit is awarded for every 12 periods. Once trainees obtain eight credits or more, they begin work on a course completion project (modifying the tools that they have mastered and applying their skills to solve problems in an actual software development environment, etc.). It takes at least a year and a half to complete the course. Classroom lectures are held to a minimum and the focus is on using tools to complete practical exercises. Trainees are required to submit a report each week. As the classes are made up of students, company employees and others from different walks of life, they are divided up into properly balanced groups to tackle problems on a small-group level.

Training in the Top SE program is focused on turning outstanding students and adults into technology leaders. In March of this year, the first graduating class of 12 students completed the program. Some of the course completion projects that were submitted by these students are already in actual use.

Course completion certificate

Yet the goal of the Top SE program is not limited to personnel training. The program also allocates some of its resources to the development of teaching materials for personnel training. "It is said that education depends on the instructor. However, to implement education in a well-balanced manner, we must not rely only on the quality of the instructors; outstanding teaching materials are also indispensable," says Professor Honiden. Eight textbooks have been completed, prepared based on approximately 400 slides per course.

NII at the Center of Information Science Education

This June, applications will be accepted for the third year of the program, which will begin in late August. In the first year of the program, almost all of the trainees were either students or persons recommended by cosponsoring companies. Starting from the second year of the program, however, the number of spaces for general entrance exam applicants was increased, and the number of students registering for the course increased as well. In this way, the training of "superarchitects" is making steady progress.

However, full-fledged development of teaching materials has not yet begun. The textbooks prepared based on lecture transcripts are distributed free of charge as a series of lecture notes. Already there have been requests for these materials from several companies and universities. How to enrich the teaching of the program at the locations where these materials are distributed is an issue that remains to be resolved. Methods being considered include providing additional explanations for the lecture notes to enable them to be used for self-study as well, and preparing a manual on the use of the teaching materials.

Moreover, Professor Honiden has an ambitious goal. "I'm thinking of starting up an educational outreach center within NII," he says. The center would develop superior teaching materials and distribute them nationwide, and then gather feedback regarding those materials in order to establish teaching methods. In this way, NII would play a central role in information science education in Japan. Naturally, the project would not be limited to Japan. NII would partner with institutions such as Carnegie Mellon University in the United States and Oxford University in Great Britain to deploy the project worldwide.

Japan's information science industry has lagged behind that in Europe and the United States. The success of the Top SE project could help Japan attain an equal footing with the other world leaders in this field.

(Written by Tomoaki Yoshito)

(*1) Top SE Project
A 5-year project for human resources development program for the creation of fundamental software, funded by FY 2004 Special Coordination Funds for Promoting Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), "SE" stands for "software engineering." The goal of the project is to create training facilities for advanced software engineers using an integrated industry-academia approach.

NII ESSAY

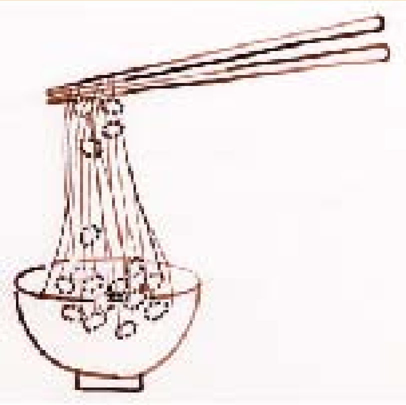
The Day that the *Natto* Vanished

Yoh'ichi Tohkura
Deputy Director - General of the National Institute of Informatics

A certain product abruptly vanishes from supermarket shelves, a strange case that one occasionally hears about. The most recent such case—the disappearance of *natto* (fermented soybeans)— received wide media coverage because it was the result of fabricated data. The fuss was caused when viewers over-reacted to a report about the dieting effects of *natto*, which had been broadcast the day before on a TV program called “Hakkutsu! Aruaru Daijiten II.” Apparently some supermarkets witnessed a doubling of *natto* sales.

The Meaning of “Information” that Changes with the Recipient

Though it is well-known that *natto* is a healthy food, there is no clear scientific proof that it aids dieting; as we have seen, the fabricated claims were discovered, and the TV station had no choice but to axe the program. The interesting point about this story is the reaction of the viewers, the way that they completely swallowed this “information,” provided by just one TV show. This phenomenon contains a number of aspects.



One is the question of people's ability to judge the veracity of information. Obviously these judgments involve knowledge, common sense and the ability to use these to analyze information. In this bizarre case of the *natto*, the attitude of the consumers who acted without any analysis or appraisal of the information they were given is noticeable. It also forces us to look at the scientific literacy of the Japanese people, which is regarded as being amongst the worst in the world.

Another aspect is the link with the trend for food faddism, in which people start to believe in and overestimate the health properties of a certain type of food. Though food faddism may be a special case, in general the way that information is accepted depends on the mentality of the recipient. In other words, the meaning of the same piece of information changes according to how seriously it is taken by the recipient. The meaning of information changes as a result of how it is amplified or dampened according to the recipient's mentality.

Requisite Ability to Analyze and Recognize Information

As we enter the information explosion society, the huge wave of information produced each day threatens to engulf us. We are already in a situation in which it is impossible to escape completely from this information. What we first of all need is an ability to discern the quality of the endlessly diverse information that this explosion had thrown up, an ability to differentiate between unproven information and information that is based on facts and is trustworthy. The ability to analyze information is what makes this feasible.

As the relationship between information and human beings becomes closer and more intricate, an ability to accept and recognize the results of our analyses in an appropriate mental state — without either over-evaluating it or under-evaluating it — is as important as our ability to analyze information. The ability to analyze information works when it is combined with a well-balanced ability to recognize it.

Surely the motivating force in managing to survive our age of ever-increasing information is an ability to analyze and recognize information, polished by a positive recognition of this patchwork data explosion — some of which is fabricated like the *natto* diet story — and our encounters with information in all its varied forms.

Weaving Information into Knowledge
NII

NII Today No. 22, July 2007 (This English language edition of NII Today corresponds to No. 36 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