# Development of the Shinshu University Online System of General Academic Resources (SOAR)

Kenji Ishizaka[1], Masashi Iwai[2,] Masato Gokan[3], Hideho Ohba[4],Ryo Sakaguchi[5]

[1,2,3,4]Shinshu University Library

[5]Media Fusion Co., Ltd.

**Abstract**

This paper discusses the development of the Shinshu University Online System of General Academic Resources (SOAR).

As a participant in the 2006-2007 Cyber Science Infrastructure (CSI) development project of the National Institute of Informatics (NII), Shinshu University is seeking to develop SOAR as an integrated academic resource system.

In addition to developing an environment for providing access to the latest academic resources within the university, SOAR is intended to promulgate university research results and research activities, both within Japan and around the world, to a broad audience. Specifically, this system achieves mutual coordination by linking e-journals and the Web of Science to the researcher directory and the institutional repository – two system cornerstones.

SOAR can be regarded as a potential model for future academic-resource systems.

Although the Institutional Repository (SOAR-IR) was developed using existing software, the Researcher Directory (SOAR-RD) is a new system based on XML technology.

**Keywords**

Cyber Science Infrastructure (CSI), institutional repository, researcher directory, XML

---

[1)] jja0101@shinshu-u.ac.jp, [2)] iwaima@shinshu-u.ac.jp, [3)] gokan@shinshu-u.ac.jp,

[4)] ohba@shinshu-u.ac.jp, [5)]sakaguchi@mediafusion.co.jp

## 1. Introduction

In recent years, electronic rather than print publication of research results such as research papers has become increasingly common in the world of academic research. At the same time, in order to publicize their research activities and to increase the visibility of individual researchers, universities around the world are aggressively seeking to establish and open to the public researcher directories, which introduce research results, research activities, and (more recently) institutional repositories – digital libraries of research results – featuring the digitized full text of research papers.[1][1][2]

In Japan, NII is promoting the establishment of institutional repositories as part of the Cyber Science Infrastructure (CSI) development project currently being implemented. Intended to develop an infrastructure for supporting e-science – advanced international and interdisciplinary research through data sharing and analysis via networks – current CSI project activities include the development of a stable, high-speed network and grid computing systems for handling large-scale processing.[2] The institutional repository is a project component intended to aid in the formation, security, and circulation of papers and other academic content. From 2006 through 2008, as part of nationwide efforts, NII is entrusting the establishment of institutional repositories to universities in Japan.[3]

However, in most cases, researcher directories and institutional repositories are separate systems. If these two systems are not linked to each other, a user interested in a paper listed in a researcher directory will need to perform another search, in an institutional repository, in order to view the full text, which is inconvenient. The same applies to a researcher arriving through some means at a paper stored in the institutional repository and wishing to find what other research the author has conducted. A system that requires considerable effort on the part of users to access information could, in a sense, be seen as counterproductive to the goal of disseminating research and researcher information.

Shinshu University has taken the opportunity presented by its participation in the NII project described above to develop the Shinshu University Online System of General Academic Resources (SOAR), an integrated academic resource system.

In addition to developing an environment for accessing the latest academic resources within the university, SOAR is intended to broadly communicate the results of research by university researchers as well as information on research activities, both within Japan and around the world. This system seeks to achieve mutual coordination by linking e-journals and the Web of Science to the researcher directory and to the institutional repository – two system cornerstones.

Since it would make the papers collected in an institutional repository easier to find using well-known search engines such as Google and Yahoo!, a system coordinating the two

---

1 The OpenDOAR site (http://www.opendoar.org/) features 928 repositories; the ROAR site (http://roar.eprints.org/) features 923 (both figures current as of August 30, 2007).

2 http://csi.nii.ac.jp/

3 http://www.nii.ac.jp/irp/index-e.html

cornerstones – the researcher directory and the institutional repository – would benefit both users and researchers. As such, SOAR represents a potential model for future academic-resource systems.

The Institutional Repository (SOAR-IR)[4] has been developed using DSpace,[5] a *de facto* standard developed by the Massachusetts Institute of Technology and Hewlett-Packard. The portions linking SOAR-IR to the Researcher Directory (SOAR-RD) and the Web of Science have been customized by AGREX INC.[6]

Since development of the Researcher Directory (SOAR-RD)[7] required XML technology, Shinshu University worked in partnership with XML database vendor Media Fusion Co., Ltd.[8]

The decision was made to open up core elements of the developed system and provide them free of charge, both in Japan and overseas, as a contribution to the academic community and to society at large.[9]

---

4 https://soar-ir.shinshu-u.ac.jp/

5 http://www.dspace.org/

6 http://www.agrex.co.jp/

7 http://soar-rd.shinshu-u.ac.jp/

8 http://www.mediafusion.co.jp/

9 SOAR information site (http://shinlis9.shinshu-u.ac.jp/soar/)

## 2. Overview of the developed system

### 2.1. Compatibility with new information environments

Since SOAR is compatible with use via the Internet through search engines such as Google and Yahoo!, it seamlessly links the Researcher Directory (SOAR-RD), the Institutional Repository (SOAR-IR), the Web of Science,[10] and collections of e-journals, making the portals transparent to users arriving from Google, Yahoo!, or other search engines. (See Fig. 1.)

The primary features of both SOAR-RD and SOAR-IR can be used from the final pages included in the search results of search engines such as Google and Yahoo! (For SOAR-RD, these pages display each researcher's profile; for SOAR-IR, these pages display summaries of each paper.)

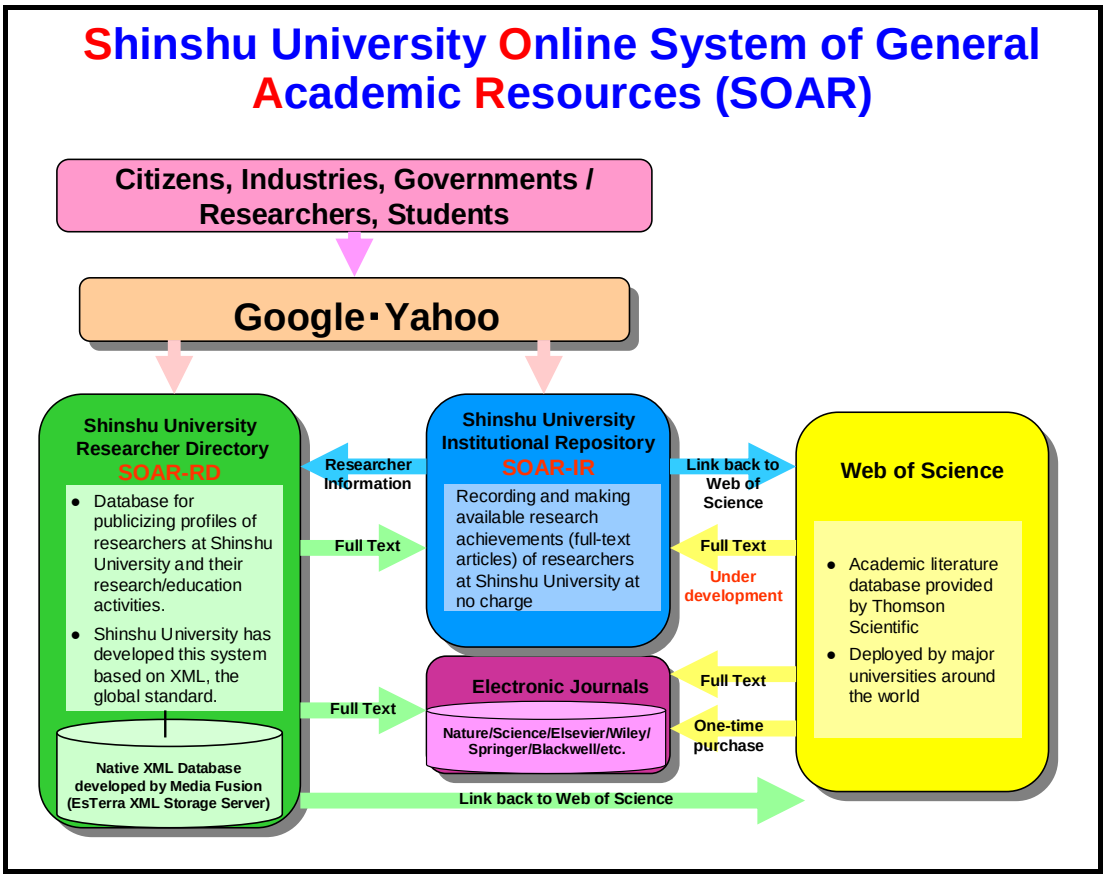


Fig. 1: SOAR overview

[10] A database provided by Thomson Scientific, used around the world as the preeminent literature and citation index database (http://scientific.thomson.com/products/wos/)

2.2. A system requiring no customization

Due to factors such as university policies and societal conditions, the items of information released to the public in the researcher directory are subject to fine-grained changes. The items of information released by researchers can also vary depending on fields of specialization and specific activities. In developing the Researcher Directory (SOAR-RD) suited to such characteristics, system designers chose to deploy XML, which offers high flexibility for redefining data structures and expansion.

Deploying XML in the Researcher Directory (SOAR-RD) search system is intended to make it possible for universities and research institutions around the world to use this system with no need for customization, since the following method of achieving system versatility within three layers (see Fig. 2) makes it possible to configure public information freely and ensures versatility, requiring no program modifications even when adding new items of data.

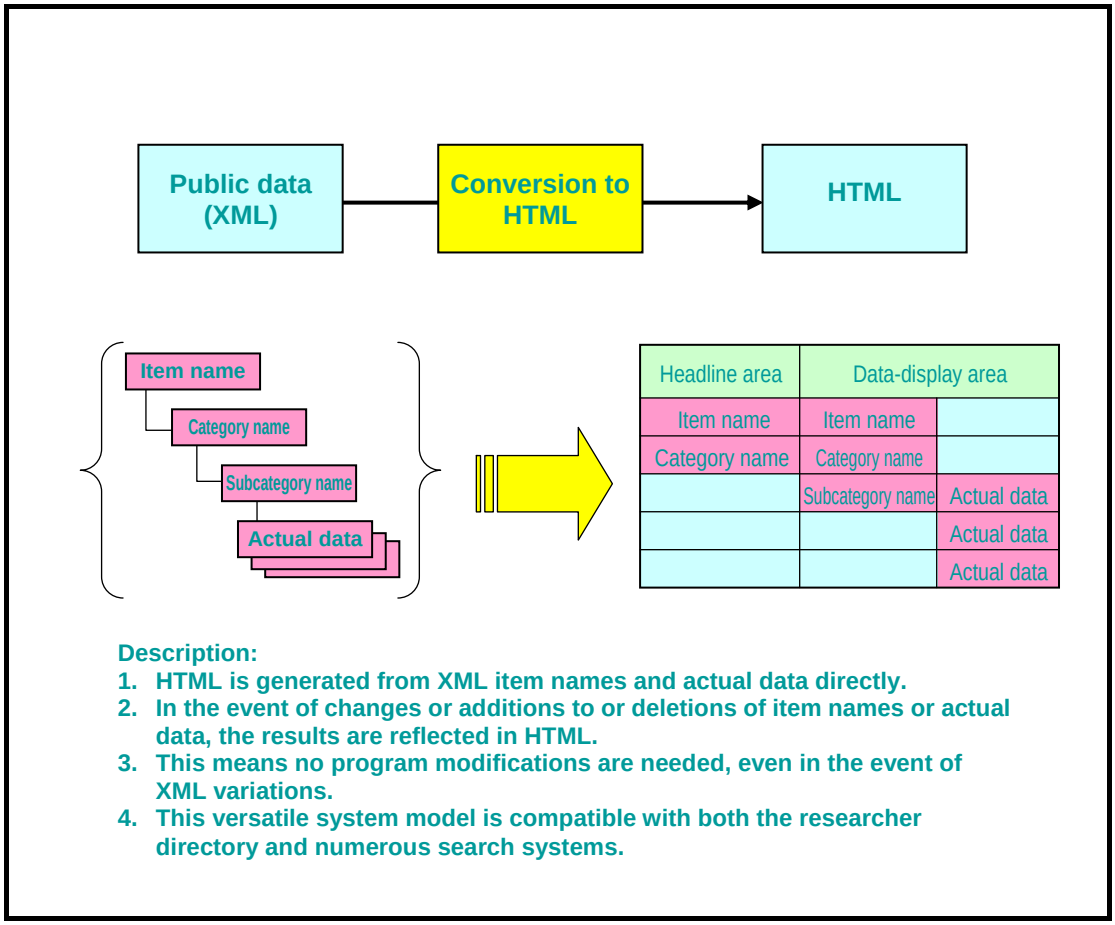


Fig. 2: Method of achieving system versatility

### 2.3. Improving the visibility of researchers and institutions

With SOAR, it's easy to move back and forth between the researcher directory and the institutional repository. The list of research results in the Researcher Directory (SOAR-RD) shows links to the text data of each paper (in the institutional repository and e-journals). Using these links, users gain smooth access to the text of literature that interests them, giving them access to a wider range of papers than they would otherwise. Reports also indicate increasing numbers of citations to open-access papers.[3] This system, which smoothly guides users to papers in the repository, can also be expected to increase numbers of citations.

In addition, a link to each author's entry in the researcher directory is displayed in each document in the Institutional Repository (SOAR-IR). Using this link, a user who has accessed a document in the institutional repository via search engines such as Google and Yahoo! can easily access information on matters such as the author's other research results and current research themes. Linking the researcher directory and the institutional repository in this way should improve the visibility of researchers and institutions by granting access to the researchers' results to a broader range of researchers from around the world.

Researchers are also free to choose whether to release their own individual items of data in the Researcher Directory (SOAR-RD). This is intended to enable researchers, who are most familiar with their own research results and the corresponding importance of these results, to tailor public access to their research results as they see fit, another end result that may lead to improvements in the visibility of researchers and institutions.

### 2.4. An orientation toward universality

SOAR will be provided free of charge to the academic community around the world. The goal is worldwide use of the system. Designed as a system that accounts for use by both researchers and the public sector, industry, ordinary citizens, and students, SOAR targets universality. Chapter Three discusses the specifics of this approach.

### 2.5. Reducing the burden on researchers

Researchers can use SOAR to carry out various types of studies through consistent, comprehensive administration of their own data. In addition, each researcher can easily update his or her own data. Chapter Four discusses related specifics.

## 3. The design of the SOAR-RD Web search page

### 3.1. Development concepts

The design of the SOAR-RD Web search page is based on the following concepts, focusing primarily on ease of understanding and ease of use:

- Minimizing the number of transitions between pages
- Enabling swift, precise identification of and access to the required information
- Featuring a function for linking to outside information sources – SOAR's greatest strength

Functions for realizing these three design concepts are described below:

### 3.2. Shortening the required steps and reducing the number of pages

When displaying profiles of researchers on the Web, the system makes it possible to use a format that divides a single researcher's profile over multiple pages by categorizing and layering data items. While this method has benefits such as limiting the volume of information displayed on a single page and organizing data through such categories and layers, it also involves the drawbacks of making it more difficult to peruse the data and increasing the number of steps required to access the necessary information.

SOAR-RD makes it easier to peruse all public data by summarizing the profile on a single page.

Efforts have also been made to minimize the steps needed by placing on the top page a text box for full-text searching and links to search results by faculty. This means profiles can be displayed through only three pages and navigating as few as two clicks – from the top page to the page listing search results, then to the page displaying the profile. (See Fig. 3.) Each page features a text box for full-text searching for rapid navigation from any page in SOAR-RD to the page listing search results and to the page displaying the profile.

The SOAR-RD Web search pages consist of four pages: the three pages described above, plus an advanced search page for designating the search targets.

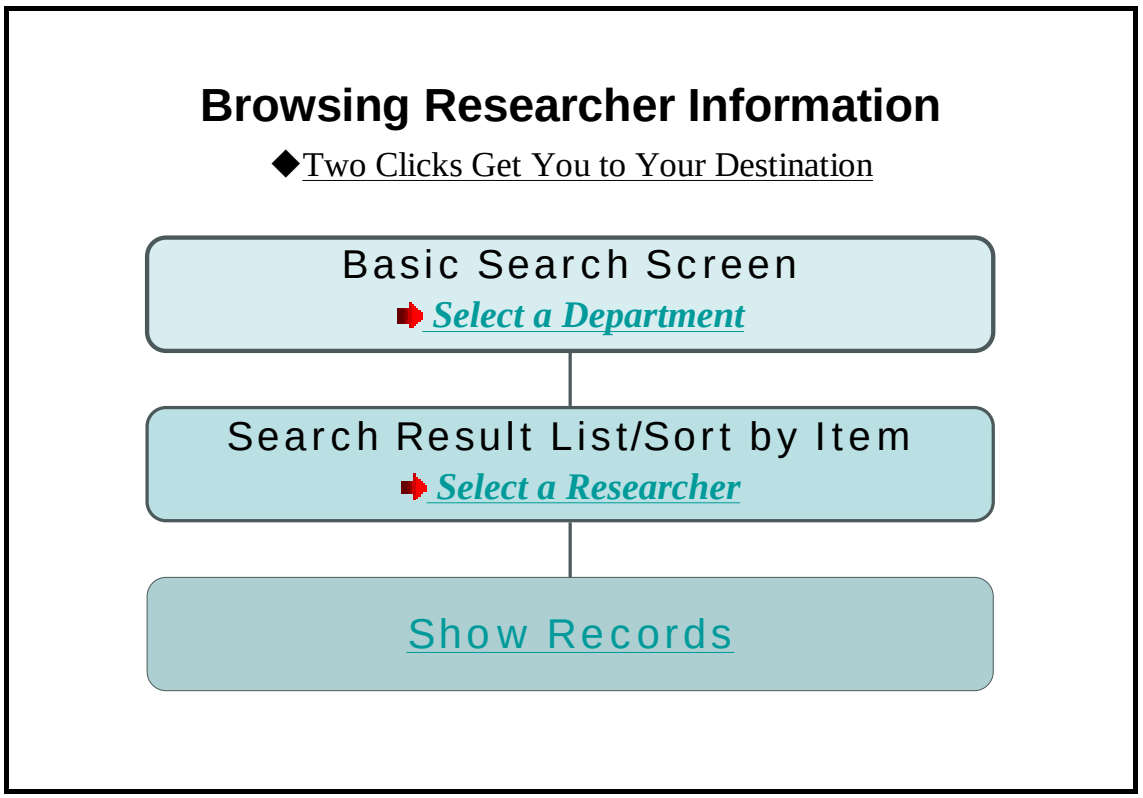


Fig. 3: Basic page transitions in SOAR-RD

### 3.3. The ability to obtain required information swiftly and precisely

#### 3.3.1. Profile page

Despite the benefits of summarizing the profile on a single page, doing so may also lead to information overload, making it difficult to identify the information needed. To counter this, the system was designed to display required information swiftly, showing topical headlines for each data item. These headlines are automatically configured based on whether input has been made for each item and whether the item has been classified as public or nonpublic. No headlines appear for empty data items or those not displayed, eliminating dead links and the display of incorrect items.

While the headlines described above are displayed on the left side of the page, the data display areas in the profile display section on the right side of the page are subdivided into two types. All public data is displayed within a single frame. To ensure ease in data perusal, basic information such as researcher names, affiliations, and contact information are displayed in a fixed position at the top, even as the user scrolls through the data section.

The data displayed includes information on years, such as academic background and Awards. To provide the latest information, such data is displayed sorted by year, in descending order. (See Fig. 4.)

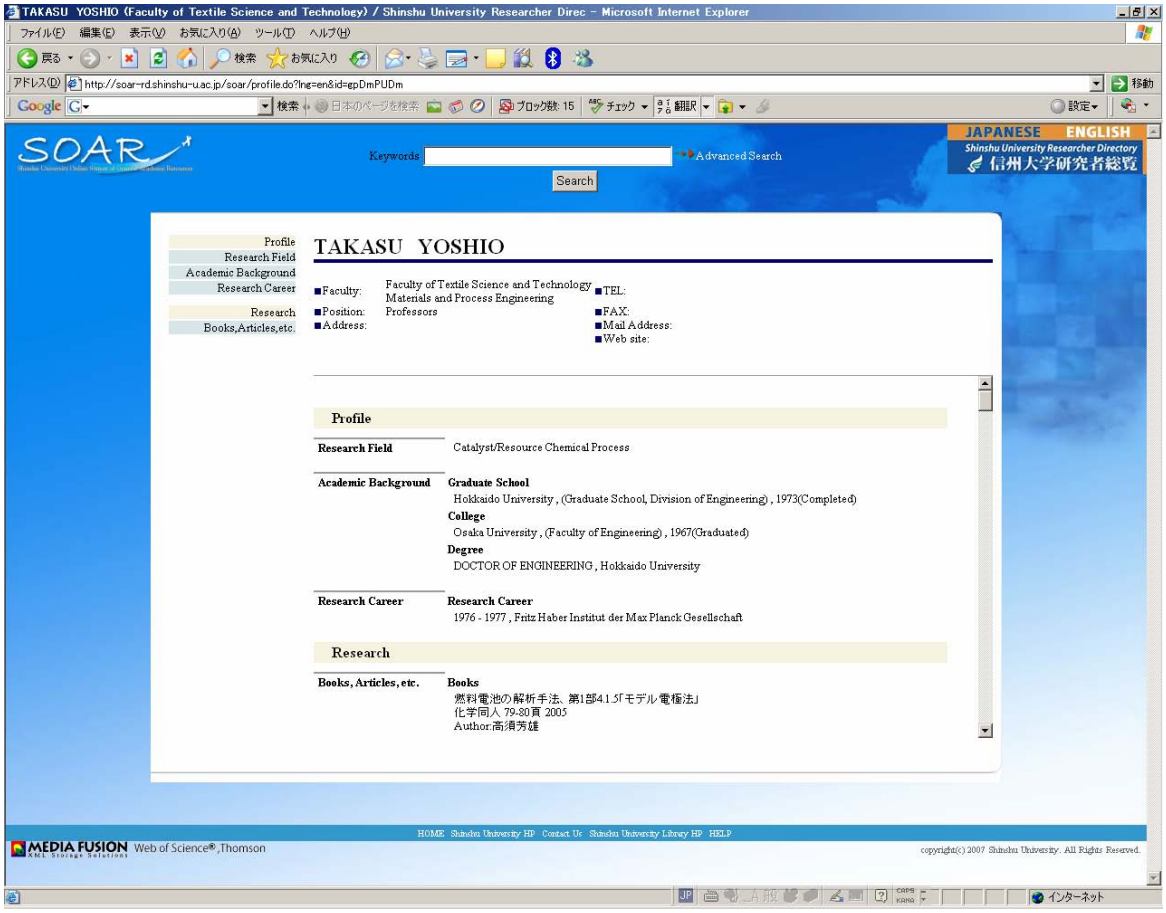


Fig. 4: Profile page

### 3.3.2. Search-results list display page

SOAR-RD's search system enables both full-text searching of papers using keywords and advanced searches designating the search target items. Since full-text searching covers all available data, such searches are expected to increase search noise. For this reason, a feature has been implemented that provides a preview of search results, displaying item titles and parts of the subject text. (See Fig. 5.) This makes it possible to identify unexpected search results when the results are listed, helping minimize usage stress attributable to search noise.

In addition, the items displayed in the list can be sorted to narrow down search results. Although each page displays 20 search results, the preview and sorting features mentioned above and the links at the top and bottom of the page should make it possible to access target information swiftly.

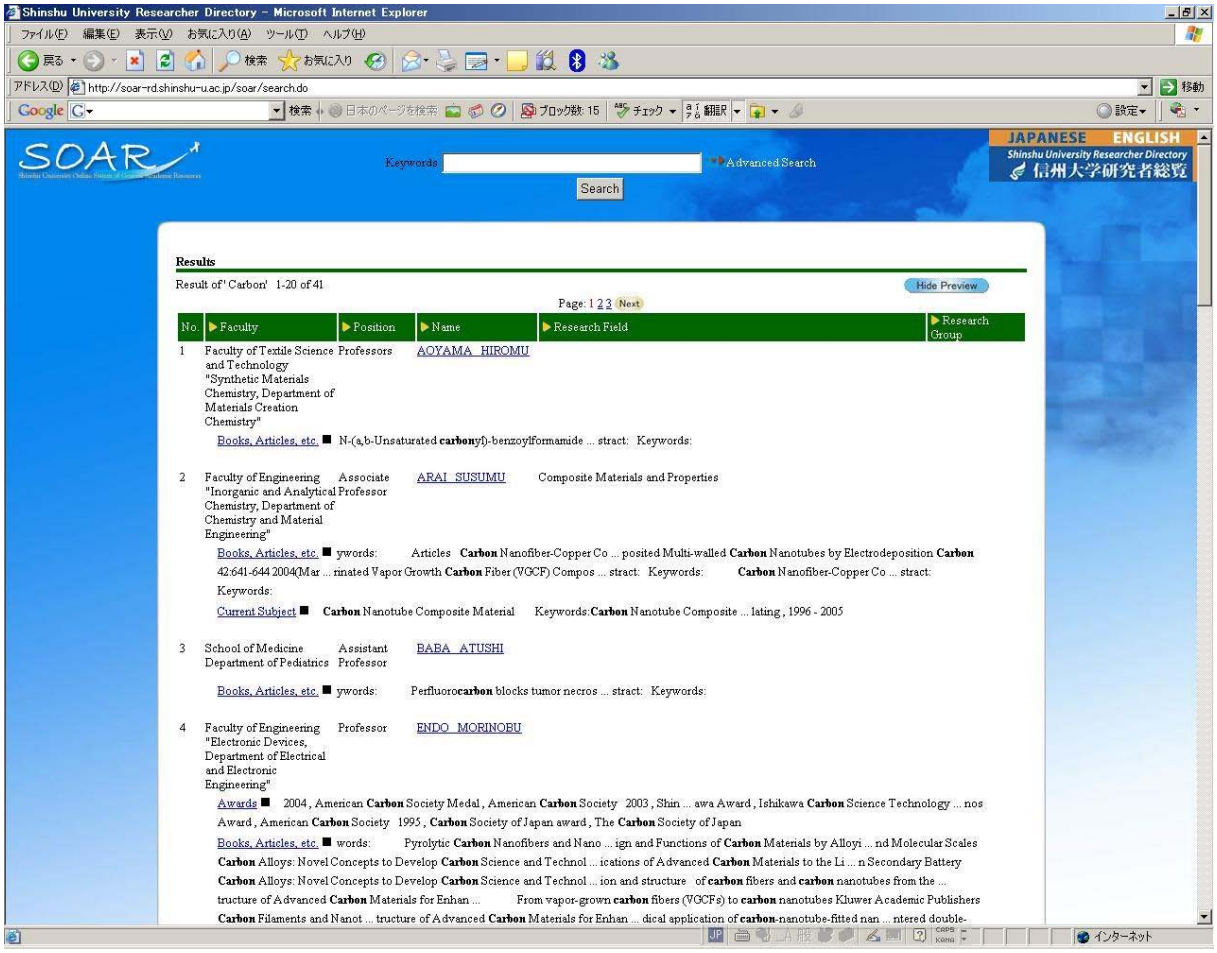


Fig. 5: Preview display of full-text search results for “carbon”

### 3.4. Links to external information sources

If the results list shown on the researcher directory does not link to other information sources, the user must take the additional step of searching another database to access the paper’s text.

Considering this additional step an impediment to the visibility of research results, in SOAR-RD, we prepared a feature that makes it possible to display for each research result links to the Institutional Repository (SOAR-ID), e-journals, and the Web of Science (see Fig. 6).

The SOAR-IR side also features links to data on the author(s) of each paper in the researcher directory. This makes it possible to reference SOAR-IR from SOAR-RD and vice versa.[11] Since anyone can access SOAR-IR, this makes it possible to publicize research results widely, even among researchers at institutions that do not subscribe to e-journals and to the general public.

[11] Reference to and from the ECS EPrints Service (http://eprints.ecs.soton.ac.uk/) of the University of Southampton School of Electronics and Computer Science’s People system (http://www.ecs.soton.ac.uk/people/) has been implemented.

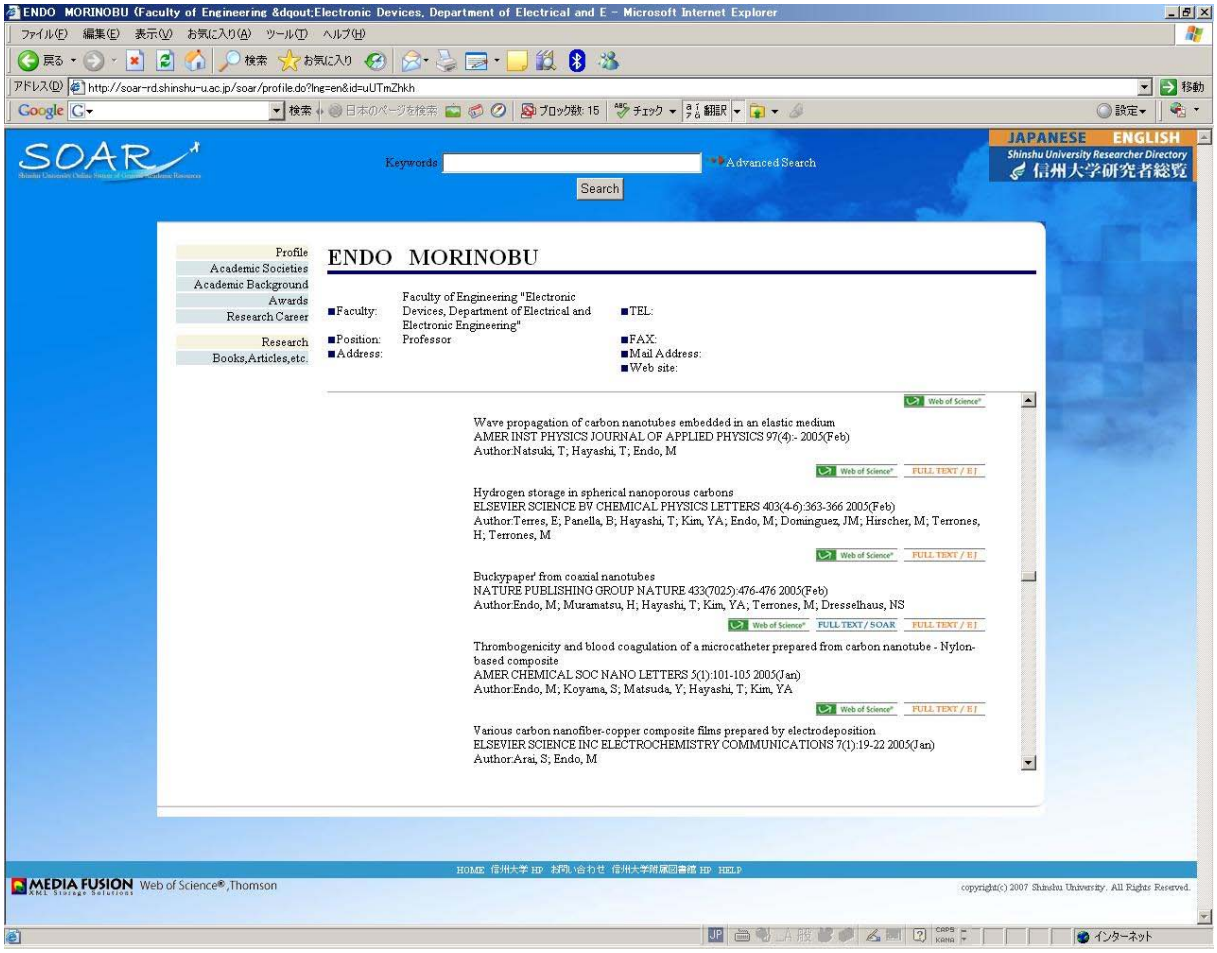


Fig. 6: Display of link buttons

Implementation of this link feature is also compatible with access not achieved via the top page (i.e., links from search engines). Users accessing the SOAR-RD profile page can obtain papers from SOAR-IR while users accessing SOAR-IR can check the profiles of paper authors and refer to their other results. In addition, information on citations in a paper and other papers citing the paper can be obtained from the Web of Science. This gives SOAR-RD – a researcher directory – many aspects of a platform linking multiple information sources.

## 4. Design of SOAR-RD's data input sections

### 4.1. Authentication system and security

When a researcher wishes to input data, the system naturally needs to authenticate the researcher's identity. We adopted LDAP authentication[12] on the university-wide portal site for authentication to consolidate IDs and passwords with those of other services, with the intention of reducing the burden on researchers.

At Shinshu University, the General Information Processing Center (GIPC) runs a lightweight directory access protocol (LDAP) server, which is used to set up the Active Campus Shinshu University (ACSU) portal site.[13] Once the researcher is authenticated by an ACSU ID and password, the user is authorized to access the personal services provided by ACSU, with no need to enter individual IDs and passwords for each service. ACSU currently provides services such as Web mail and scheduling.

As part of this service, we incorporated the above ASCU functionality into an updating function, so that researchers can update certain data, as needed. Specifically, individual user information and passwords are centrally managed on the LDAP server, and only a correspondence table for LDAP user IDs and researcher-directory researcher IDs is kept on the researcher-directory side.

This enables a massive increase in user convenience, since researchers do not need new IDs and passwords and the system itself requires little administration about the authentication. The standard LDAP protocol also lowers obstacles to adoption by other users, including other universities.

Both the ACSU login screen and this system's data-update screen implement encrypted communications using the secure sockets layer (SSL) protocol to ensure the security of data paths, allowing researchers to administer their own data from both on campus and off campus.

---

[12] The technical standard for LDAP authentication is defined in RFC4513 [4].

[13] https://acsu.shinshu-u.ac.jp/

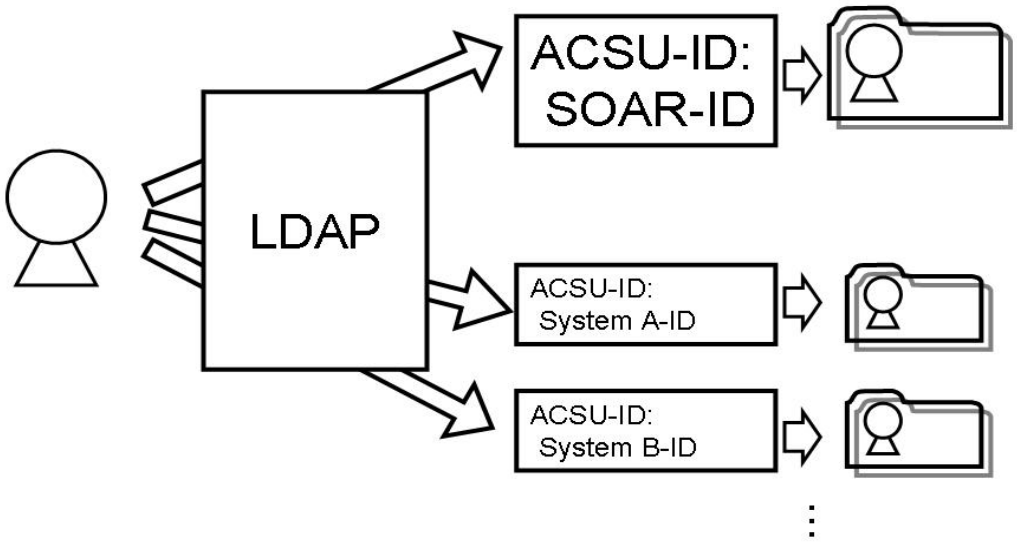


Fig. 7: Conceptual diagram of LDAP authentication

## 4.2. Data input methods

### 4.2.1. Use of Microsoft Excel files

In many cases, entries to databases on researcher information are made by inputting items one at a time on a webpage. [5][6]

While this method involves an interface (the Web) available to anyone and presents a relatively low user threshold, it also has the following drawbacks:

- It involves large numbers of pages, making it difficult for a user to check his or her own data as a whole
- When the user stops inputting data before he or she is finished, the data is released to the public in an incomplete state. A single series of input operations requires time to ensure all data is properly input.
- Since it uses large numbers of pages, the system entails significant system development and maintenance costs and time requirements.

To resolve these drawbacks, this system adopts a method whereby all data on a single researcher is loaded using a single file in Microsoft Excel format. The data is checked and updated through

downloading and uploading of this Excel file between the user and the master server.

This approach resolves the drawbacks described above. A user can check his or her data as a whole at any time on a local computer, since the data is stored in a single file. The user can also input data at any time, even entering data piecemeal and uploading it only when complete. Finally, the only page required for input is a file upload/download page.

The Excel format was adopted for two reasons: Excel is considered the most commonly used spreadsheet application,[14] which many researchers already have running on their computers and for the most part already know how to use; doing so allows use of CabineX, an existing software product from Media Fusion, which was in charge of developing this feature. CabineX converts Excel files to XML data.

### 4.2.2. Excel-file structure

Each Excel file consists of a separate worksheet for each data category, such as academic background, career history, research results, and educational achievements, with data-entry fields arranged in tables. In each table, columns are data items (fields) and rows are individual records (see Fig. 8). A researcher can add to or revise the data at any time while checking his or her own data as a whole.

[14] According to Tanaka et al. [7], Excel's share of the spreadsheet market reached 80% in 2000. Although new players have since appeared, including Open Office, the basic pattern appears unchanged.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 研究活動業績3（著書、発表論文等） | | | | | | | |
| 2 | | | | | | | | |
| 3 | | 日本語版研究者総覧チェックフラグ | 英語版研究者総覧チェックフラグ | ReaDチェックフラグ | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | | | | | ・研究者総覧における非公開項目は | | |
| 7 | | | | | 研究業績（著書、発表論文等）＊ | | | |
| 8 | | 選択 | 選択 | 選択 | 著書又は発表論文の標題＊又は発表テーマ＊ | 出版社＊又は学会主催団体 | 掲載誌名＊又は会議名＊ | 巻＊ |
| 9 | 1 | | | | | | | |
| 10 | 2 | | | | | | | |
| 11 | 3 | | | | | | | |
| 12 | 4 | | | | | | | |
| 13 | 5 | | | | | | | |
| 14 | 6 | | | | | | | |
| 15 | 7 | | | | | | | |
| 16 | 8 | | | | | | | |
| 17 | 9 | | | | | | | |
| 18 | 10 | | | | | | | |
| 19 | 11 | | | | | | | |
| 20 | 12 | | | | | | | |
| 21 | 13 | | | | | | | |

Fig. 8: Sample Excel file

Each row represents one data record. Flags to indicate whether to display each record on the Web are entered under columns B, C, and D. Data items are added from column E.

Embedded in the Excel file is hidden data on which cell stores which tag of the XML data. (See Chapter 5 for details.) Even though the system uses an XML database, researchers need not be aware of the XML. All they need to do is enter data in accordance with the tables in Excel.

### 4.2.3. Processing through on-screen display

When this file is uploaded to the server, it is converted to XML data by the CabineX Office2XML Module, then released on the Web after conversion to HTML data.[15] The program monitors file uploads at all times, and the conversions are automatic. In addition, uploaded files are stored on the master server.

A public/nonpublic flag is affixed to each individual record on the Web (see columns B - D on the left side of Fig. 8). This gives the researcher precise control over whether specific information is available to the public.[16] The system skips records with nonpublic flags when it generates HTML files, allowing researchers to input data they do not wish to release to the public but wish to record in the system for research purposes or as personal notes.

---

[15] Implemented using JavaServer Pages (JSP) files

[16] Each item included within a record (in Fig. 8, “publisher,” “publication,” etc.) is configured in the system in advance as either public or nonpublic. For example, information not suitable for public release, such as that on pending patents, can be controlled by setting all such information as nonpublic.

Internally, the system uses UTF-8 character encoding, allowing it to handle a wide range of research fields, including those that require use of special characters.

See Chapter 5 for details of the technical aspects of the system covered in this section.

### 4.3. Handling surveys

To reduce burdens on researchers – one of the system's goals – we have implemented a feature that makes it possible to provide data to the Directory Database of Research and Development Activities (ReaD) and to output information as needed for various documents, based on data recorded in the system. Once data has been recorded to the system, there is no need to re-input the data for various surveys.

ReaD (http://read.jst.go.jp/) is a site operated by the Japan Science and Technology Agency (JST) to summarize and release to the public data on the career histories and research results of researchers throughout Japan. ReaD data can be updated in two ways: through online updates, in which individual researchers update their own data and through data exchange, in which research institutions jointly collect and update in bulk data on affiliated researchers. This system enables automatic handling of this data exchange to allow automatic data provision with no effort on the part of researchers. As with the decision on whether to make data public or nonpublic, researchers control whether to provide individual data records to ReaD in their Excel files.

At the development stage, the system's handling of various types of surveys was first subject to the planning protocols for government grants-in-aid for scientific research. We also envision that the system will handle applications and reports for various other types of subsidies, as well as plans submitted to university accreditation committees.

Government grants-in-aid for scientific research represent the leading competitive research subsidies in Japan, budgeted at 190 billion yen per year and provided to 55,000 research projects annually. Most university researchers apply for these each year,[17] and the planning protocols for such applications require the provision of information on past research results and grants received in connection with the research theme covered by the application. Since researchers need to collect appropriate information from their own research results and provide it within application forms, preparing the protocols involves significant effort. Since other research surveys require similar effort, the associated burden on researchers to date has been considerable. The feature described here is intended to reduce this burden.

This feature makes it possible for researchers to choose from the recorded data the records they wish to output. The content researchers wish to cover is expected to vary, depending on where the documents are submitted and on their research themes. As when downloading Excel files, researchers will log into the system using LDAP authentication and select the desired document

[17] As announced by the Ministry of Education, Culture, Sports, Science and Technology [8], 116,243 applications were received in the 2006 fiscal year – approximately 0.68 per university professor. In some cases, multiple researchers may be involved in a single research project, or a single researcher may have submitted multiple applications.

format and records to be output. The system outputs the selected records in accordance with the items of data and the selected format. (These formats must be loaded into the system in advance.)

### 4.4. Handling changes to data items

Given various factors, including the possibility that outside organizations may change the items required in surveys as well as internal university needs, the data items in the researcher directory may change from time to time. For example, in the ReaD site described under section 4.3, data items and description methods are revised occasionally.[18] It's also possible that new data items will need to be added when implementing new links to other systems (such as course syllabi and e-learning contents).

In the event of such changes to data items, the native XML database implemented by this system requires only small-scale changes to the data, requiring no large-scale data restructuring. This is one of the benefits of using an XML database.

However, the Excel files must handle tasks such as changes to table and worksheet structures and to tag information used for XML conversion. Procedures for migrating existing data to new Excel files are also needed.

### 4.5. Items for achieving coordination with external resources

To realize coordination with external resources as described under section 3.4, the research results data includes fields for the input of each of the resources linked to. Numbers from the Handle System[19] of the Corporation for National Research Initiatives (CNRI) are employed for links to the institutional repository. For links to e-journals, DOI[20] are employed and ISI unique article identifiers[21] for Web of Science use. Link buttons are displayed when HTML files are generated only if a value has been input to the relevant field.

---

18 Information on the latest changes can be found at http://read.jst.go.jp/info.html

19 A system for assigning permanent, dedicated addresses to items included in repositories (http://www.handle.net/)

20 An abbreviation for Digital Object Identifier, a unique address assigned to a digital file (http://dx.doi.org/)

21 Unique ID numbers assigned to records in the Web of Science

## 5. Technical aspects of SOAR-RD development

### 5.1. SOAR-RD system overview

Fig. 9 shows a flowchart for this system, from data input by the researcher through user searching and browsing.

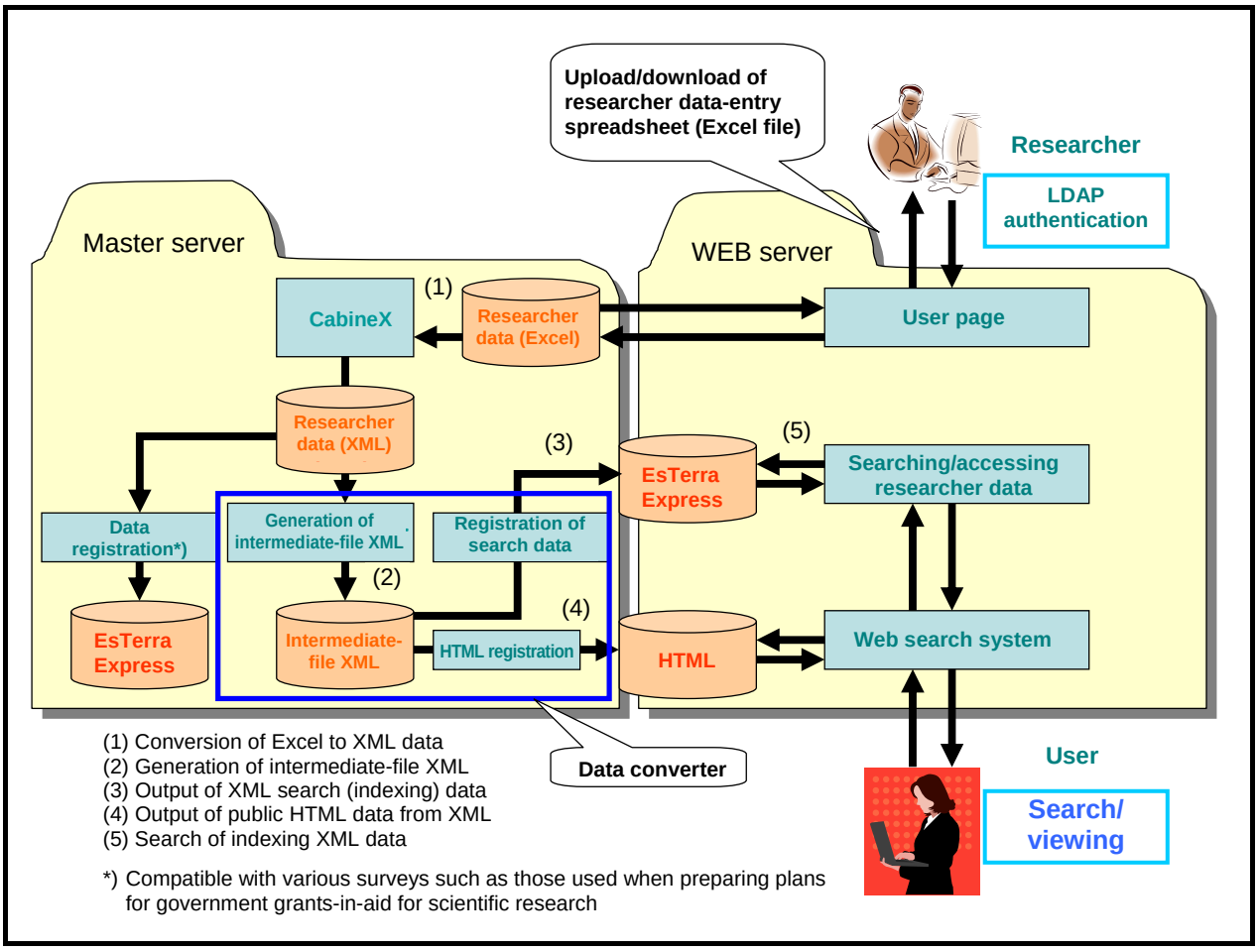


Fig. 9: SOAR-RD system overview

The following sections describe the technical aspects of the following characteristics of SOAR-RD system development: 1) implementation of Excel files into a single workbook; 2) implementation of public data; and 3) achieving high-speed full-text searching.

## 5.2. Implementation of Excel files in a single workbook

### 5.2.1. Relation between Excel and XML

Ordinarily, in converting data for an Excel file to XML, a single worksheet is prepared for data input and another to specify conversion definitions.[22] For this reason, for multiple worksheets, one Excel file must be prepared for each data worksheet. But the limitations of Microsoft Office prevent multithreading with multiple Excel files, forcing less efficient sequential processing.

### 5.2.2. Implementation using CabineX

This system is implemented using CabineX, which can fetch the value of each item in an Excel file and output it in Extensible Markup Language (XML) format, suitable for secondary use.

The CabineX application automatically converts Microsoft Word and Excel files prepared using the CabineX Template Maker to XML or saves them as flat files using the CabineX Office2XML Module. CabineX XML2Office Module can also be used to export XML to Word or Excel files.

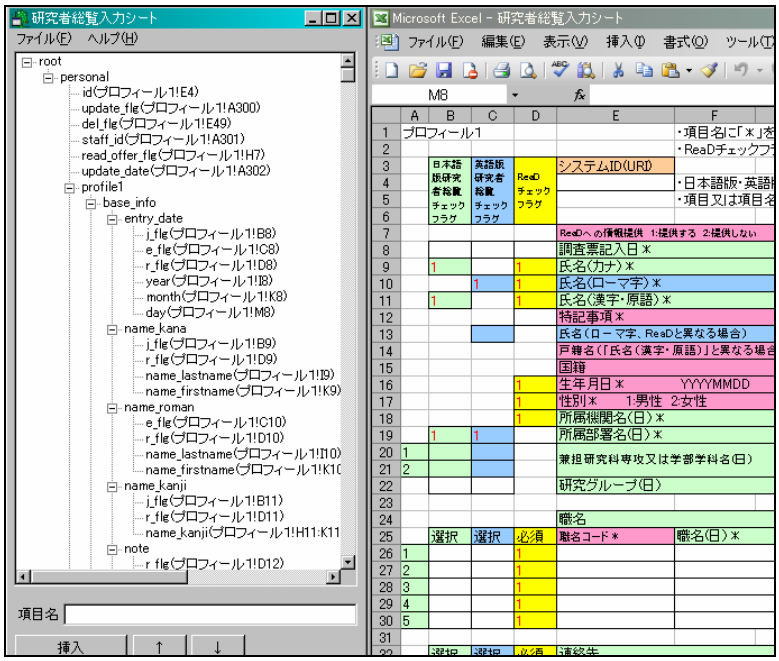


Fig. 10: Associating items using TemplateMaker

A Microsoft Excel spreadsheet is shown to the right-hand side of the screen.
On the left-hand side is a TemplateMaker screen showing the XML structure. This can be used to designate the Excel cells linked to each XML tag.

[22] Definitions of mapping between Excel cells and XML elements

#### 5.2.3. Using a single workbook for Excel worksheets

One characteristic of CabineX is its capacity to hold definitions for conversions between Excel files and XML within Excel files. In CabineX, conversion definitions are held as properties,[23] not specified in a worksheet, making it possible to specify conversion definitions for multiple worksheets. This characteristic makes it possible to store researcher-directory data consisting of multiple Excel worksheets in a single workbook.

### 5.3. Implementation of public data

#### 5.3.1. Data converter structure

The converter program converts the XML data converted in CabineX to data for external release. The converter handles the following two tasks:

- Generating XML, extracting only data intended for public release
- Generating HTML and the indexing XML needed for searching (i.e., generation of public data) from extracted XML

Since the XML extracted from CabineX includes all data input to the Excel file, it also includes data that need not be released to the public. For this reason, processing before generating data for public release generates intermediate XML (intermediate-file XML) consisting of only public data. Certain defined processing[24] is carried out at the same time.

Indexing XML, intended specially for searching, is generated based on intermediate-file XML. This XML has a structure optimized solely for high-speed full-text searching, for which ordinary native XML databases are poorly suited. HTML for public release is concurrently generated from the intermediate-file XML. All public data is generated using this processing. See Section 5.4 for details of indexing XML.

---

[23] Application is currently pending for a patent examination of data-generation algorithms (as of August 2007).

[24] Data processing required for sorting and display

### 5.3.2. Data and converter versatility

In comparison with structured data formats, XML offers much higher multiplicity and flexibility. In general, converting data entails fixing data formats before and after conversion. This is because links in ordinary data conversion are made by designating paths or by using names for reciprocal conversion. For this reason, schema need to be redefined when adding data fields to structured data. There is no need for such redefinition of schema with XML.[25] This system inherits the data-structure flexibility characteristic of XML; even if the XML structure of an intermediate file is changed, there is no need to change the converter program itself.[26]

## 5.4. Achieving high-speed full-text searching

### 5.4.1. Using an XML database

For the following reasons, this system stores data internally using XML:

- XML is an open data format adopted as a global standard
- XML is ideal for data structure changes
- XML makes it easy to achieve layered structures
- XML has a high level of affinity for Web systems
- XML data is in text format

When handling large volumes of XML data as with this system, storing data as flat files cannot provide adequate search response speeds. For this reason, some kind of database needs to be used in the implementation.

If the data structure were fixed, the most effective implementation might involve a relational database (RDB). However, implementation using a RDB poses difficulties for a system like this one that must handle changes to the data structure. For this reason, we decided to implement this system using the EsTerra XML-only database engine. EsTerra is a native XML database engine that makes it possible to store and search XML documents in their native tree structure, enabling high-speed searching of data in layered structures – a challenge for RDBs.

### 5.4.2. Indexing XML

The searchable XML data itself was designed to increase search speeds.

The actual XML data has a versatile XML structure suited to secondary use. While its structure is ideal for extracting data from items and outputting this data to other forms and for generating

---

[25] For other than well-formed XML, DTD and other definitions are required.

[26] XML structures can be changed using TemplateMaker, in accordance with modification rules. Style sheets and structural associations require redefinition.

Web pages, it is less than optimal for data searching. And since all data is included, whether public or nonpublic, storing this data on a public server would pose security-related issues.

For these reasons, this system maintains separate XML data specially for indexing purposes (indexing XML). This indexing XML consists solely of publicly releasable data, which poses no security-related issues, while enabling high-speed full-text searching using the XML database.

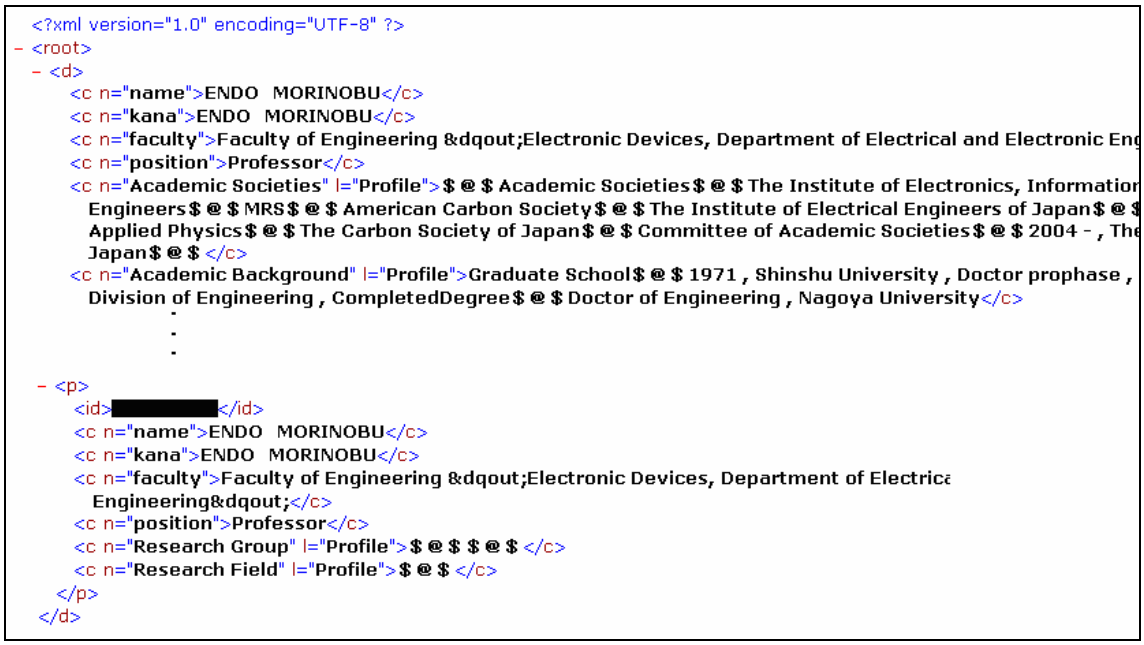


Fig. 11: Sample of indexing XML

The indexing XML uses the XML structure shown in Fig. 11. Fragmentated into individual researcher, this data stores public information in child elements by category. The category items shown in the left-hand pane of Fig. 4 are specially structured for searching, storing category names, item names, and actual data. In addition, data fetching for display is optimized by storing in a separate fragments only the items for display (the “p” nodes in Fig. 11).

This structure is effective for high-speed partial-match query processing of statements like the following:

/root/d[c/text()=~“Ryo Sakaguchi”]

Internally, when inputs can be made to multiple fields, these text strings are administered as a single text node connected using the separator “$@$” rather than administering them as separate elements. This increases the speed of full-text searches or partial-match query processing while reducing the number of blocks used. Subqueries such as “p/id” can be

executed on the “d” elements fetched using the above query format to fetch the ID of the relevant researcher.

## 6. Conclusion

SOAR was developed during the 2006 fiscal year, with public beta testing for the institutional repository undertaken from January 2007 and for the researcher directory from March 2007. During testing, we migrated data from existing databases and made system adjustments. The formal release came in August 2007. This system development was chosen by NII as one of the best examples of subcontracted projects under the CSI project for the 2006 fiscal year.

From the development stage through public beta testing, explanatory meetings for researchers were held to solicit questions and perspectives from them. Researchers raised many queries on matters related to operation and the system, including the scope of results that can be registered and copyright handling. Instruction manuals prepared beforehand appear to have clarified most aspects of Excel file input.

One possible course of action for future system development is in the area of coordination with other systems. At present, the only function implemented for coordination with external systems is between the repository. However, coordination with systems for course syllabuses, e-learning, and research expenses, among other uses, should also be possible.

Core system elements are currently provided free of charge, and we are ready to provide the system free of charge to universities and research institutions wishing to use it for researcher-directory purposes. Such institutions are expected to develop additional functions independently.

We hope that this system will contribute to the scholarly communication among universities and research institutions.

## Acknowledgements

Shinshu University participates in Area I: Start Up and Enhancement Projects of the NII Institutional Repositories Program 2006-2007, part of the National Institute of Informatics’ Cyber Science Infrastructure (CSI), from which it receives funding.

We wish to express our gratitude for the cooperation and support of the following individuals:

From the Shinshu University Library: Director General, Prof. Akio Nomura, Mr. Ei’ichi Momoi, Mr. Hisamori Tezuka, and Mrs. Chiyo Ogiwara.

From Media Fusion Co., Ltd.: President Atsushi Sakakibara, Mr. Koan Kurata, and Mr. Akira Murata.

**References**


[1] Gerard van Westrienen and Clifford A. Lynch, “Academic Institutional Repositories -- Deployment Status in 13 Nations as of Mid 2005,” D-Lib Magazine, Volume 11, Number 9, 2005. (online) available from http://www.dlib.org//dlib/september05/westrienen/09westrienen.html3

[2] H. Itshumura, “The expansion of institutional repositories in Japan -- Changes in scholarly communication and accumulation,” Current Awareness, No. 291, pp. 12-15, 2007. (Japanese)

[3] S. Harnad and T. Brody, “Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals,” D-Lib Magazine, Volume 10, Number 6, 2004. (online) available from http://www.dlib.org/dlib/june04/harnad/06harnad.html

[4] R. Harrison Ed., “Lightweight Directory Access Protocol (LDAP): Authentication Methods and Security Mechanisms,” Request for Comments 4513, 2006. (online), available from http://www.ietf.org/rfc/rfc4513.txt, (accessed July 31, 2007)

[5] T. Oie, T. Ueta, Y. Ochi and Y. Yano, “Education and research database in Tokushima University – Development and operation of the database orienting periodical publication and information disclosure --,” Research in University Evaluation, No. 3, pp. 31-50, 2003.

[6] S. Nima and M. Kimura, “Web Data Administration System using cooperated Server Model – Application to the electric page of KENKYUSHA SOURAN –,” Bulletin of Kyushu Women's University, Natural sciences, Vol. 39, No. 4, pp. 27-39, 2003.

[7] T. Tanaka, T. Yasaki and R. Murakami, “Economic analysis of network externality – One idea on competition policy under externality,” Collaborative research report CR 01-03, Competition Policy Research Center, 2003. (Japanese) (online), available from http://www.jftc.go.jp/cprc/reports/cr0103.pdf, (accessed July 26, 2007)

[8] Ministry of Education, Culture, Sports, Science and Technology, “Distribution of FY2006 Grants-in-Aid for Scientific Research,” 2006. (online) available from http://www.mext.go.jp/b_menu/houdou/18/10/06092713.htm