Researcher Name Resolver: identifier management system for Japanese researchers

Int J Digit Libr (2014) 14:39–58
DOI 10.1007/s00799-014-0109-z

Researcher Name Resolver: identifier management system for Japanese researchers

Kei Kurakawa · Hideaki Takeda · Masao Takaku · Akiko Aizawa · Ryo Shiozaki · Shun Morimoto · Hideki Uchijima

Received: 30 December 2012 / Revised: 22 January 2014 / Accepted: 22 January 2014 / Published online: 19 February 2014

Abstract We built a researcher identifier management system called the Researcher Name Resolver (RNR) to assist with the name disambiguation of authors in digital libraries on the Web. RNR, which is designed to cover all researchers in Japan, is a Web-oriented service that can be openly connected with external scholastic systems. We expect it to be widely used for enriched scholarly communications. In this paper, we first outline the conceptual framework of RNR, which is jointly focused on researcher identifier management and Web resource linking. We based our researcher identifier scheme on the reuse of multiple sets of existing researcher identifiers belonging to the Japanese grant database KAKEN and the researcher directory ReaD & Researchmap. Researcher identifiers are associated by direct links to related resources on the Web through a combination of methods, including descriptive mapping, focused crawling on campus directories and researcher identification by matching names and affiliations. Second, we discuss our implementation of RNR based on this framework. Researcher identifiers construct uniform resource identifiers to show Web pages that describe researcher profiles and provide links to related external resources. We have adapted Web-friendly technologies—e.g., OpenSearch and the RDFs of Linked Data technology—in this implementation to provide Web-friendly services. Third, we discuss our application of RNR to a name disambiguation task for the search portal of the Japanese Institutional Repositories Online to determine how well the researcher identifier management system cooperates with external systems. Finally, we discuss lessons learned from the entire project as well as the future development directions we intend to take.

Keywords Researcher identifier · Author identifier · Name disambiguation · Author search · Digital library · Identifier management · Researcher Web resource · Web resource linking · Linked data · Institutional repository

K. Kurakawa (✉) · H. Takeda · A. Aizawa · R. Shiozaki
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan
e-mail: kurakawa@nii.ac.jp

H. Takeda
e-mail: takeda@nii.ac.jp

A. Aizawa
e-mail: aizawa@nii.ac.jp

R. Shiozaki
e-mail: shiozaki@nii.ac.jp

M. Takaku
National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan
e-mail: takaku.masao@nims.go.jp

S. Morimoto
Kanazawa University Library, Kakuma-machi, Kanazawa, Ishikawa 920-1192, Japan
e-mail: morikun@adm.kanazawa-u.ac.jp

H. Uchijima
Tsukuba University Library, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
e-mail: huchijim@tulips.tsukuba.ac.jp

1 Introduction

The use of digital libraries has been increasingly spreading in the academic world, and more and more research articles and academic books are being archived on the Web. In particular, researchers are accessing more and more electronic articles, papers, and books on the Web. As researchers ourselves, we frequently use general and academic search engines to look for research papers, books, academic software, and data. We

often browse the directories of publishers' digital libraries to observe trends in a given field. We then browse the knowledge network of a field by following links on the Web, which might take us to citation networks, relevant sources, or increasingly social networks.

However, we often encounter difficulty when trying to distinguish who the authors of such content are. We cannot classify papers by exact author name, e.g., "John Smith", because in many cases different authors have the same first and last names. In terms of aggregating papers associated with the same authors, the names printed on the papers should not be the only evidence used because an author may change his or her name (e.g., after marriage), or the names may appear in different formats (e.g., a given name may be shortened to only an initial). Publishers individually regulate these formats, which adds to the confusion.

Name disambiguation of authors is clearly recognized as necessary for enriching scholarly communications [1–3]. A variety of stakeholders can get benefits from this. Readers can retrieve exact search results by author and search for research development paths by researcher, available range of research fields, and collaboration capabilities over different disciplines. Research and development managers can track the productivity of the researchers in their organizations.

Two approaches are generally taken with the name disambiguation of authors: clustering by machine and identifier assignment. Clustering by machine is essentially automatic and therefore effective when dealing with large volumes of bibliographic data. However, it cannot perform complete name disambiguation without any errors. In the identifier assignment approach, a list of published authors is manually created. An identifier is assigned to an author and can be used as a key to clarify the relevance to an author's publications. This approach can accomplish complete name disambiguation without any errors, but it is much more time-consuming and expensive.

Some digital library systems have started using the clustering approach to aid in author searches. In these cases, the bibliographic metadata of papers are automatically clustered by means of computational algorithms. Thomson Reuters implemented a "Distinct Author Identification System"[1] in the Web of Science, while Elsevier implemented "Scopus Author Identifiers"[2] in Scopus. Microsoft Academic Search,[3] CERN's high-energy physics digital library, INSPIRE,[4] and the computer science-specific digital library CiteSeer X[5] use clustering with its author searches. The Japanese digital library of the National Institute of Informatics (NII), CiNii Articles,[6] takes the same approach with its author search function.

Identifier assignment is another name disambiguation approach, and it can be applied to the above digital library systems along with the clustering computation. Thomson Reuters provides a researcher identifier assignment service, ResearcherID,[7] in which individual researchers sign up and claim their publications. These publication claims are publicly displayed on the site and linked to the bibliographic record on the Web of Science.

Identifier assignment can also provide additional value for scholarly communications. In the case of researchers, they can easily organize self-publications and automatically produce curriculum vitae, while for the administrators of research societies, the manuscript-tracking system can help them keep track of author activities and communicate with exterior financial bodies. For the administrators of organizations and grant-funding agencies, they can easily access the research productivity portfolios of researchers in whom they are interested.

In this paper, we present our Researcher Name Resolver (RNR),[8] an identifier management system designed to cover all researchers in Japan. It is a Web-oriented service that can be openly connected with external scholarly systems. We are confident that its wide use will enrich scholarly communications.

In the next section, we provide an overview of related works dealing with author identifiers. We then describe the conceptual design of RNR in the following section. Its design consists of the researcher identifier framework, which is the core concept of identifier management, and of Web resource linking, for the added-value function. This conceptual design is fundamental for implementing RNR. After that, we discuss the results of our application of RNR to a name disambiguation task for the search portal of the Japanese Institutional Repositories Online (JAIRO).[9] We then conclude the paper with a brief discussion of this study and a mention of the future directions of RNR development.

2 Related work

When we started our project in 2008, name disambiguation of authors was still an unsolved problem, and a variety of stakeholders tried to deal with technological development

[1] http://images.webofknowledge.com/WOK46/help/WOS/h_da_sets.html.

[2] http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authoridentifier.

[3] http://academic.research.microsoft.com.

[4] http://inspirehep.net.

[5] http://citeseerx.ist.psu.edu/index.

[6] http://ci.nii.ac.jp.

[7] http://www.researcherid.com.

[8] http://rns.nii.ac.jp.

[9] http://jairo.nii.ac.jp.

and struggle with business strategies to achieve name disambiguation of authors for their relevant databases. The production release of the Open Researcher and Contributor ID (ORCID)[10] in 2012 was thought as the end of the struggle. But, in fact, it was a starting point for collaboration among such the stakeholders. The name disambiguation of authors is still being a big challenge.

In this section, we give a historical overview of author identifiers and the relevant systems.[11] These systems are not enough for production level because they lack sufficient disambiguation accuracy and full coverage of researchers. The following describes library catalogs, subject repositories, institutional repositories, national practices of researcher identifier, ORCID, and more general creator identifier, ISNI.

Technically speaking, library catalogs have been using author identifiers since the early 1980s, which was the dawn of online computing services. Most library cataloging systems have a name authority file to collocate bibliographies for the exact same author. Author identifiers are usually represented in alphabetical and numerical character formats and written in bibliographic metadata along with author name. Bibliographies and name authority files are in a position to reference each other.

The Library of Congress has been making the name authority file of its cataloging system available to the public on the Web[12] since 2002. Most national libraries give anyone access to their catalogs and name authority files, including the National Diet Library of Japan[13] whose service has started at 2011. Recently, many library catalogs have become more Web-friendly in terms of search and retrieval processes. The Virtual International Authority File (VIAF)[14] is an international project to assign author identifiers to identify the same authors in the name authority files of national libraries throughout the world [4,5]. The VIAF prototype system opened to the public in 2007.

Subject repositories have been extensively used by researchers in various domains since the 1990s, which is when the Web began to emerge and evolve. The most well-known subject repository is arXiv,[15] a preprint server of the mathematics, physics, and computer science domains, established in 1991 [6]. arXiv expanded in 2005 to store name authority files within the system and was reformed in 2009 as a service called "Author Identifiers" [7]. In this service policy, authors are optionally assigned an identifier, and they can sign up with the system and obtain their identifier while they are in the process of submitting a paper or any time after. The publication list of an author can be displayed on Facebook by referencing the author identifier. In the economics field, RePEc[16] is the main subject repository [8], and it has also provided an author identifier service (called "Author Service") since 1999. This identifier service was extended to cover all research fields in 2008 and is now known as AuthorClaim.[17] Researchers sign up with the system to obtain an author identifier and claim their publications. The data registered in this system are all public under the CC0 license of the Creative Commons.

Institutional repositories are a set of services offered by a university that disseminates digital materials created by the institution and its members [9]. They are one of a range of responses to what is generally known as the serials pricing crisis from the end of 1990s. Institutional repositories empower the open access movement [10]. Even though the institutional repositories are a result of open access policy, in practical terms, for users to be able to read all the articles they contain, name disambiguation for authors is recognized as necessary [11].

One of major repository software applications, DSpace, added an authority control function in version 1.6.0 in May 2010. An item metadata composed of several fields is revised so that each field has an identifier attribute. Users can input any characters as identifier in the identifier attribute. Default installation gives an assistant module to put an identifier through referencing the identifiers of the Library of Congress authority file on the Web. In 2007, another repository software application, E-Prints, added a function to hold an authority file in version 3.0. Users prepare this authority file to list authors' names and e-mail addresses. When a user inputs part of an author name in the metadata field during the item registration process, the E-Prints system suggests author names and embeds the relevant e-mail addresses. These identifiers make it possible to perform precise author searches in DSpace and E-Prints.

The Names Project[18] in the UK, from an early age of repository software development, aims at being the name authority for UK repositories. Author identifiers are generated by author clustering from the bibliographic records of the British Library's Zetoc service by means of computational algorithms [12]. The Names Project has provided the service since 2008 and subsequently created a plug-in for the E-Prints auto-completer in 2011 that gets access to the API of the Names server. In this case, a person uniform resource identifier (URI) assigned by the Names server is fulfilled in the identifier field.

[10] http://orcid.org.

[11] Comparative studies conducted by Fenner [25] and Hill [26] are informative for readers.

[12] http://authorities.loc.gov.

[13] http://id.ndl.go.jp/auth/ndla.

[14] http://viaf.org.

[15] http://arxiv.org.

[16] http://repec.org.

[17] http://authorclaim.org.

[18] http://names.mimas.ac.uk/advanced-search.

Author identifiers are assigned by national organizations that wish to manage the achievement lists of researchers. The SURF foundation in the Netherlands assigns a Digital Author Identifier (DAI) to all the faculty, researchers, and staff of their national universities and research institutions. NARCIS,[19] which is the Netherlands' national research portal for information about researchers and their work, comprises author pages arranged using DAIs and listing-related publications [13]. Dutch institutional repositories store publication metadata with DAI, which has been harvested by NARCIS since 2006. In Brazil, the National Council for Scientific and Technological Development (CNPq) maintains a curriculum vitae and institutions database called the Lattes Platform.[20] All working researchers as well as Ph.D. and Master students in Brazil are assigned a Lattes Curriculum ID and must report their achievements for inclusion in the Lattes Curriculum, a component of the Lattes Platform [14].

Intermediate and catalytic identifiers are currently being assigned for authors and are poised for use in a variety of scholarly systems. The ORCID is a promising example, especially as it is not limited to one country but rather assigned to researchers all over the world. The ORCID organization was founded in the US in 2010 as a non-profit entity and includes major academic publishers, academic societies, universities, research institutions, and funding agencies among its key members. The ORCID system is based on the ResearcherID system designed by Thomson Reuters and can be interconnected with external systems, including ResearcherID, Scopus, VIVO,[21] INSPIRE, RePEc, Author Resolver of ProQuest, and PubMed, among others.

More abstract and comprehensive identifiers for authors are standardized at ISO 27729:2012 as part of the International Standard Name Identifier (ISNI). ISNI assigns identifiers for the creators of books, music, movies, articles, and any other creative object. It functions as a bridge identifier to maintain connections among public identities [15].

3 Researcher identifier framework

Identifier representation is the first and central consideration for building an identifier management system. To prepare for a set of identifiers covering all researchers in Japan, it seems obvious that building it from scratch would take a very long time and be very expensive. The best way to minimize time and cost is to take advantage of researcher identifiers that are already in use.

Based on this premise, we designed a researcher identifier scheme, researcher identifier management process, and its URI form. The researcher identifier is associated with a researcher profile to identify the researcher. We also designed a researcher profile data scheme and researcher profile data management process.

3.1 Identifier scheme

For the background to consider identifier scheme of researchers in Japan, here we mention how many researchers are in Japan. The Ministry of International Affairs and Communications (MIC) conducts the Survey of Research and Development every year and reports the number of researchers in Japan who are doing research work for a specific theme. As of 2012, the number of researches in Japan was 844,430. Of these, 490,920 researchers belong to business enterprises, 39,598 researchers belong to non-profit institutions and public organizations, and 313,912 researchers belong to universities and colleges. University and college researchers are composed of 187,730 faculty staffs, 72,079 medical staffs and others, and 70,991 doctor course students.

To cover all the currently active researchers in Japan, we utilized the research grant awards database KAKEN[22] (operated by NII) and the researcher directory ReaD & Researchmap[23] (operated by the Japan Science and Technology Agency (JST) and NII). These two databases cover major part of researchers in Japan, including university faculties, staff, students, and research institution staff. The important thing is to cover all researchers involved with tax-funded research.

KAKEN is the awards database of the Grants-in-Aid for Scientific Research (namely KAKENHI) administrated by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Society for the Promotion of Science (JSPS). These grants, which cover all research fields and all stages of research activities in Japan, date from 1965 and are the only nationwide grants in Japan, not to mention among the biggest of Japanese research budget items. Most researchers in universities and colleges in Japan have applied for this grant and are registered in the system. For example, 68,406 research project proposals were accepted in 2012 (its acceptance rate was 51.7 %), and of which 74 (0.1 %) projects were conducted in business enterprises, 5,306 (7.8 %) projects were conducted in non-profit institutions and public organizations, and 63,026 (92.1 %) projects were conducted in universities and colleges. Totally, 202,341 researchers are registered in the system. In recent 5 years (from 2008 until 2012), it covers 111,840 registrants, which indicates the coverage of currently active researchers.

[19] http://www.narcis.nl.

[20] http://lattes.cnpq.br.

[21] http://vivoweb.org.

[22] http://kaken.nii.ac.jp.

[23] http://researchmap.jp.

Table 1 Identifier scheme of the Researcher Name Resolver

ID (13-digit number)	Description
10000xxxxxxxx	ID assignment for KAKEN registrants. xxxxxxxx is fulfilled with the KAKEN researcher number (8-digit)
200xxxxxxxxx	ID assignment for ReaD & Researchmap registrants xxxxxxxxxx is fulfilled with the ReaD researcher code (10-digit)
3xxxxxxxxxxx	ID assignment for registrants not affiliated with KAKEN or ReaD & Researchmap

ReaD & Researchmap is a researcher directory in Japan that evolved into a social network infrastructure among researchers in 2011. The researcher directory originated from a "Survey on Academic Research Activities" conducted by MEXT's predecessor, the Ministry of Education, Science and Culture, in 1961, and was compiled as a database by NII's predecessor, NACSIS, in 1990. JST constructed an original database of researchers starting in 1998 and combined both databases in 2002. About 220,000 researchers are registered in the current directory. The number of researchers who are simultaneously registered in both ReaD & Researchmap and KAKEN is about 80,000.

The two databases differ in that KAKEN covers only staff members connected to universities and research institutions, while ReaD & Researchmap additionally covers some Master course and doctoral students and is free to register. In addition, the two databases have slightly different identifier schemes. In KAKEN, a researcher is identified by a unique 8-digit number called a "Researcher Number", while in ReaD & Researchmap, a 10-digit number called a "ReaD Researcher Code" is used. This code is used to upload batches of researcher profile information to the backend system and is usually unnoticed at the bottom of the researcher profile pages.

Our identifier scheme uses both of these identifiers to register researchers, which results in the pump-priming effect for the rest of researchers who are not registered to register themselves. Table 1 shows the identifier scheme used in the Researcher Name Resolver; it consists of a 13-digit number. For the KAKEN researcher numbers, a string of "10000" is padded for the prefix of the number. For the ReaD researcher codes, a string of "200" is padded for the prefix of the number. For those who are not registered in either KAKEN or ReaD & Researchmap, we assign a prefix of "3".

3.2 Identifier management

Our identifier scheme will potentially create multiple assignments for one researcher. For example, the same researcher may be sequentially assigned a "3", "200", and "10000" prefix identifier at various intervals. This happens because we use two kinds of external identifiers, which results in the reservation of two external identifier-related spaces. The first advantage of the proposed scheme is its ability to provide a higher coverage of researchers. The second advantage is that if the external identifier is already known, the translation of identifiers between the two systems is quite easy, even if done manually. If you know the KAKEN researcher number "12345678", it means that you already know the identifier "100012345678" for the same person in the Researcher Name Resolver.

To solve this multiple assignment problem, we must deal with the relations among the identifiers of the same researcher and let just one identifier be representative among them for reference. To decide which one should be the representative, we prioritize identifiers: the "10000" prefix identifier is the first priority, the "200" is the second, and the "3" is the third. We chose this order for the practical reason that active researchers have the Grants-in-Aid for Scientific Research researcher number that is well known.

Formally, we assume a researcher $r_n \in R$, where R is the set of researchers. r_n might have multiple identifiers, as in $id^{(r_n)} \in ID^{(r_n)}$, where $ID^{(r_n)}$ is a set of identifiers of r_n. If we decide the priority of an identifier by priority$(id^{(r_n)})$, the representative identifier of r_n is determined by

$$\text{representative\_id}(r_n) := \underset{id^{(r_n)} \in ID^{(r_n)}}{\arg\max} \ \text{priority}(id^{(r_n)}).$$

In this case, the identifier is classified as $id_{C1} \in C1$, $id_{C2} \in C2$, $id_{C3} \in C3$, where $C1$ is the "10000" prefix identifier class, $C2$ is the "200", and $C3$ is the "3". The priority is $\text{priority}(id_{C1}) \succ \text{priority}(id_{C2}) \succ \text{priority}(id_{C3})$.

This representative framework must be adapted the same way within a classified set of identifiers because our identifier scheme depends on external identifier schemes that are out of our control, and in practical operation, multiple identifier assignment happens so frequently that it cannot be ignored. For example, once researchers are assigned their Grants-in-Aid for Scientific Research researcher number, they typically keep it for life. Nevertheless, some researchers are assigned a researcher number twice or more due to clerical errors. When this happens, the most recently assigned number is used for further applications and the other assigned numbers become obsolete. Even so, the old documents are still archived with the obsolete numbers, and those numbers remain valid in the archives.

3.3 URI representation and transfer

We assume that the identifier will be referenced on the Web. According to the WWW architecture [16], the WWW is

an information space in which items of interest, referred to as resources, are identified by global identifiers called URIs. A URI consisting of an identifier and an RNR prefix represents a researcher entity. If an identifier is "1000012345678", the URI of the researcher is http://rns.nii.ac.jp/nr/1000012345678, which leads to the personal page of the researcher.

The WWW architecture recommends that "a URI owner SHOULD NOT associate arbitrarily different URIs with the same resource". The fact that our identifier scheme allows multiple assignments is in conflict with this recommendation. The problem is that different URIs might identify the same researcher. To prevent this, the WWW architecture suggests that "an agent that receives a URI SHOULD refer to the associated resource using the same URI, character-by-character". Therefore, we use the redirection mechanism of a Web server so that the incoming URI of an identifier will be redirected to the URI of the representative identifier.

3.4 Profile data scheme

At a glance, the profile data of researchers seem not to be an essential element of the identifier management system. However, researcher profiles are indeed necessary, for two reasons. The first reason is that, to make a researcher distinct from others, we require sufficient academic attributes of the person in question to the extent that we can identify his or her personality. Curriculum vitae is a good example of such identification of an academic personality. The second reason is that, to further develop the functionality and enable linkage from a researcher page in the system to an external researcher page on the Web, we require at least evidential information in order to identify the researcher, no matter how a personal page will be processed by a computational algorithm.

As stated previously, RNR deals with researcher profile data from KAKEN. The KAKEN researcher profile has the Grants-in-Aid for Scientific Research researcher number as its key attribute and consists of the name and affiliation of the researcher, a list of grant award IDs, a list of research fields, and an up-to-date list of research keywords relevant to the researcher. The name of the researcher is represented in three ways: Kanji (Chinese characters), Katakana (Japanese phonetic characters), and Romaji (English transliteration).

RNR also deals with ResearcherID-based profile data. We think of the ResearcherID provided by Thomson Reuters as a distinctly international researcher directory, so we expanded its XML schema for profile data uploading to include the Japanese language and academic cultural differences. This expanded profile data schema additionally lists external researcher identifiers in which both external researcher identifier and service name are described in a set of defined fields.

3.5 Profile data management

KAKEN researcher profile data and ResearcherID-based profile data need to be uploaded to the system. A KAKEN researcher profile is inevitably linked to a "10000" prefix identifier by its researcher number. When a ResearcherID-based profile includes a "10000" prefix identifier, both profiles are linked together through the identifier. Otherwise, a list of external researcher identifiers described in the profile is checked for the researcher, in which a group of identifiers for the researcher is updated. The profile is linked to the representative identifier of the group.

Formally, we assume a researcher $r_n \in R$, where r_n might have multiple identifiers $id^{(r_n)} \in ID^{(r_n)}$, and external identifiers $extid^{(r_n)} \in EXTID^{(r_n)}$ where $EXTID^{(r_n)}$ is a set of external identifiers of r_n. The profile has a list of external identifiers, so an RNR identifier is linked to external identifiers $link(id^{(r_n)}, extid^{(r_n)})$, and if another researcher $r_m \in R$ has an external identifier $extid^{(r_m)} \in EXTID^{(r_m)}$ that is equivalent to an $extid^{(r_n)}$, it derives a unified group of identifies for the researcher, as

$$
\begin{aligned}
&G_1 = \{id^{(r_m)}, extid^{(r_m)}\},\ G_2 = \{id^{(r_n)}, extid^{(r_n)}\},\\
&\qquad \exists extid^{(r_m)} \wedge \exists extid^{(r_n)} \wedge extid^{(r_m)} = extid^{(r_n)},\\
&\rightarrow\ G_3 = G_1 + G_2 = \{id^{(r_m)}, extid^{(r_m)},\ id^{(r_n)}, extid^{(r_n)}\}
\end{aligned}
$$

where G_1, G_2, and G_3 are groups of identifiers.

In this framework, the last document archived in a researcher's timeline is typically the one chosen to be displayed, and it takes priority over the others. If some profiles are deleted, it reconfigures a timeline and may change how the profile is displayed.

4 Web resource linking

An additional advantage of RNR is its ability to link to external Web resources. To find resources for a researcher, we often use free search engines (e.g. Google, Yahoo, or Bing) and put the researcher's name and some keywords to obtain relevant search results. However, we found that these results include both relevant and irrelevant documents on the Web, and we often cannot verify whether the names highlighted in the snippets actually belong to the researcher in question. Direct links to the researcher resources are more valuable than search links with researcher names because the researcher entity resolution quality of the former is higher than that of the latter.

We take two approaches to create direct links: descriptive mapping and focused crawling with automated mapping.

4.1 Descriptive mapping

4.1.1 Implicit identifier reuse

An RNR identifier is mapped from the Grants-in-Aid for Scientific Research researcher number and the ReaD researcher code. If either of these systems has a Web service to redirect a user request with the researcher number or code to the relevant researcher page, the RNR can have a direct link to the page.

KAKEN and CiNii Articles provide an access method via the Grants-in-Aid for Scientific Research researcher number that RNR can use to create direct links. Currently ReaD & Researchmap does not provide any access and so RNR cannot create direct links in the same way.

4.1.2 ID table batch load

Another way of creating a direct link is to use a mapping table to link different external identifiers related to the same researcher. Uploaded profile data might contain a list of external researcher identifiers and external service names, which can then be used to specify the URL format for accessing the service. Direct links are created so as to consist of an identifier and the URL format. The profile data are supposed to list identifiers for several external services, e.g., ResearcherID, ORCID, KAKEN, ReaD, and NACSIS-CAT (a Japanese university holdings catalog of books). It also lists researcher page URLs in a campus directory as direct links.

As in the backend system, RNR regularly imports the identifier table that lists the Grants-in-Aid for Scientific Research researcher number, the ReaD researcher code, the ReaD & Researchmap account name, and J-GLOBAL[24] ID for each researcher, which is provided by JST. J-GLOBAL is a service for browsing academic information resources attributed to JST. Direct links for these services are then made from this table.

4.2 Focused crawling and automated mapping

4.2.1 The target: campus directories

A slightly more difficult way of making links to external researcher resources is to crawl campus directories and make links between campus directory researcher pages and RNR researcher pages.

In Japan, as of 2012, we have 783 universities, 372 junior colleges, and 57 colleges of technology, and of 783 universities, 86 are national universities, 92 are public universities, and 605 are private universities, which was reported in a School Basic Survey conducted by MEXT.

In these schools, it has been mandatory for campus directories to be open to the public under the Ordinance for Enforcement of School Education Act[25] since 2011. Campus directories are required to include basic faculty information and achievement descriptions related to research and education. This information is arranged for each individual related to the university such that the faculty description looks like a general curriculum vitae along with some distinguished attributes. The user interface of a campus directory to search faculties also varies depending on how it fits into the organization structure.

4.2.2 Site structure sensitive crawl

We designed and implemented a crawler specifically to address the issue of crawling a variety of campus directories. Our crawler searches for researcher pages over the linking network and collects a complete set of the URLs of a campus directory. In the interest of optimizing time and cost, we chose to use a kind of focused crawling method.

Focused crawling, which was initially introduced by Chakrabarti et al. [17], is an approach in which the crawler seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively narrow segment of the Web. In the past decade, several types of focused crawlers have been proposed, including focused crawling with reinforcement learning, context-focused crawling, intelligent Web-forum crawling, focused crawling in a Web database, and crawling focused on geographic locations [18]. These crawlers learn by training on topics related to the general task, after which they decide which links they should crawl. Our crawler, in contrast, behaves under a set of rules for each campus directory. This rule-based technique is simpler and more robust and feasible than the others because campus directories have generic and simple structures in common.

We implemented our crawler by customizing Nutch, which is an open source Web crawler [19]. The components and functions of our crawler are shown in Fig. 1. We implemented plug-in modules to extend the components filled with a background color of the standard Nutch crawling framework. The crawling sequences are listed below.

1. The injector reads the seed URL—that is, a list of URLs to start crawling from—and then registers these URLs in the CrawlDB.
2. The generator scores the URLs in the CrawlDB in accordance with its customization and then extracts the "top N" URLs as the URL list to fetch.
3. The fetcher references the URL list and fetches the contents from the Web. The fetcher detects whether a URL points to a general HTML page or a JavaScript page. If it

[24] http://jglobal.jst.go.jp.

[25] http://law.e-gov.go.jp/htmldata/S22/S22F03501000011.html.

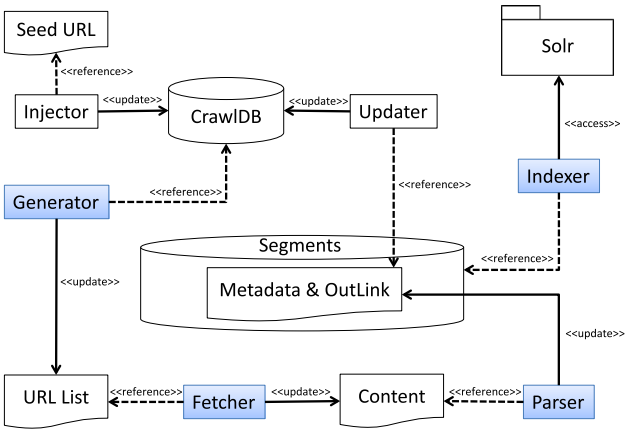

Fig. 1 Nutch-based crawling workflow

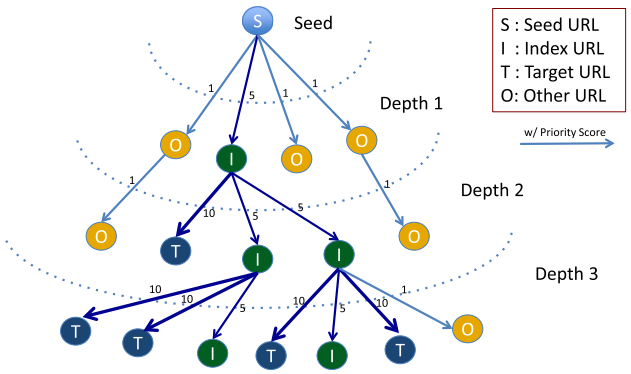

Fig. 2 Focused crawling path

is a JavaScript page, the fetcher triggers a browser engine to translate it into a general HTML page.

4. The parser parses the contents to extract the relevant text and URLs for each campus directory in accordance with the set of rules for that campus directory and then stores the text as metadata and the URLs as the out-link in the segments.
5. The updater adds the out-link URLs to the CrawlDB.
6. For the number of depths to crawl, repeat steps 2–5.
7. The indexer references and indexes the metadata in the segments for the search engine package "Solr".

For each university campus directory, we set seed URLs to start crawling and define regular expression patterns of URLs for campus directory index pages and researcher pages. The generator gives 10 points for researcher pages, 5 points for index pages, and 1 point for other pages; in other words, researcher pages are the highest priority. We also define scraping patterns for the parser by means of XPath and regular expressions. The parser scrapes the researcher pages to build a list of names, including Kanji name, Katakana name, English name, affiliation, job title, and amendment date.

The focused crawling paths our crawler goes through depth-by-depth are shown in Fig. 2. For one fetch cycle, the generator extracts the top 50 URLs based on the priority scores and the fetcher fetches their contents. We applied our crawler to the campus directories of Niigata University and show its exemplary series of fetch cycles in Fig. 3. The types of URLs the crawler fetched in each cycle are depicted. The crawler retrieved index pages and some other pages in the early fetch cycles and then successively retrieved researcher pages because the index pages list the researcher page URLs. After the crawler retrieved all the index pages and researcher pages in the CrawlDB, it retrieved other pages for a while and then the fetch cycle terminated. The terminate condition is that the crawler exceeds 20 % of fetch cycles devoted to retrieving only other pages.

To determine how well our method can collect all researcher pages of a campus directory, human testers compared the number of crawled researcher pages with the number of registered researchers in the campus directory. We set 226 university campus directories (49 national universities, 25 public universities, 148 private universities, and 4 junior colleges) as the crawling target and ultimately retrieved 86,395 researcher pages. The number of crawled researcher pages and the number of registered researchers were the same for all campus directories.

4.2.3 Naïve-automated mapping

To create links between RNR researcher pages and campus directory researcher pages, we first have to identify the researchers associated with these pages. Of course, if we just compare the researcher names commonly described in the researcher pages, we encounter the name ambiguity problem. To tackle this problem, we focus on affiliations to decrease the probability of the same family name and personal name: i.e., a case of the same family name and personal name can be resolved by consulting the affiliations. This is not a complete solution for the name ambiguity problem, but it is a first step and therefore useful to apply. We estimated the number of researchers in an affiliation and how many researchers share the same family name and personal name (see Appendix). The number of researchers in an affiliation ranges from approximately 100 to 5,000. In an affiliation of such a size, the probability of the same family name and personal name—in other words, a non-unique name—is 0.0–0.6 % (zero to 30 researchers).

Family name distribution has long been of interest to biological researchers and physicians for its role in the biological and sociological aspects of human nature. Baek et al. [20] report a comprehensive family name distribution study in which the family name distribution is expressed as an approximate formula $P(k) \sim k^{-\gamma}$, where k is the size of the family (in other words, the number of individuals

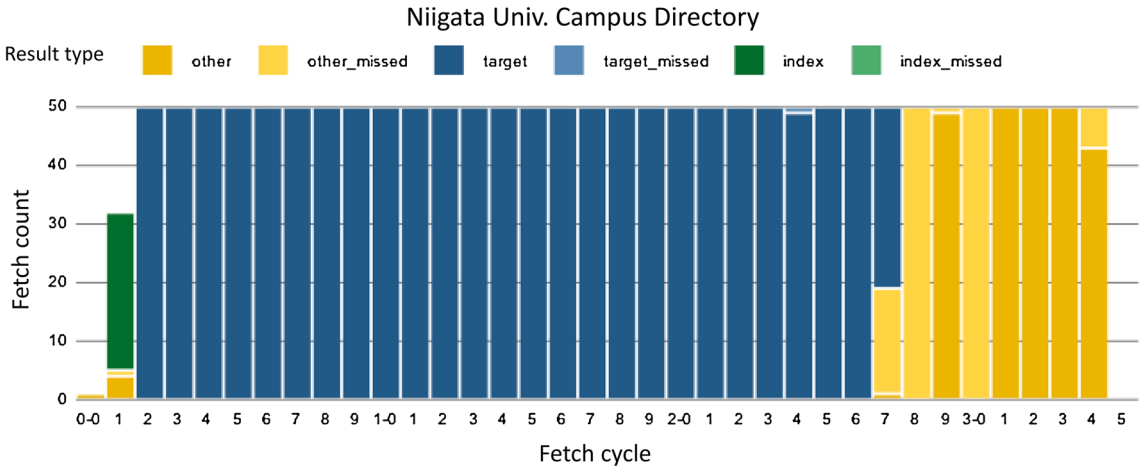

Fig. 3 Fetching priority URLs in fetch cycles, colored by URL type and access result

who have the same family name) and γ is a constant. The number of observed family names N_f increases either logarithmically or algebraically when the sampled population size N increases. They consider two groups of distributions: those that draw on power-law form[26] and those that draw on exponential Zipf's plot (exponential scale/rank). These two groups are caused by the appearance of new family names and therefore depend on cultural behavior and social dynamics. In a Japanese case, the family name distribution $P(k)$ is also shown in a power-law formula $k^{-\gamma}$, where the exponent $\gamma \approx 1.75$ and $N_f = N^{0.65}$ [21]. Family names in the United States and in Berlin also display power-law behavior with a similar exponent $\gamma \approx 2.0$ [22]. In contrast, a Korean family name distribution shows different behavior, where the exponent $\gamma \approx 1.0$ and $N_f = \ln N$ [23]. The number of family names in Korea grows in a smaller step when the population size increases than that of other countries because Koreans consider inventing new family names as taboo. Koreans totally accept branching out using new regional origins (but with the same family name) in Korean culture. They use the birthplace of the family names as additional information, i.e., where the ancestor of the family came from, to distinguish families.

Full name distribution is of interest in the current study because we need to know full name distributions in order for the proposed system to work. However, there are few studies on full name distribution for us to reference. Family name distribution has been studied well than full name distribution. This might be because family name distribution can seem like a natural phenomenon since parents cannot control the family name of their child.

To illustrate a full name distribution of a Japanese name, we accessed the JAPAN/MARC[27] name authority file and counted the number of same family names and personal names. JAPAN/MARC is the bibliographic catalog of holdings in the National Diet Library of Japan and covers both publications from Japan and from throughout the world that are written in Japanese. It includes books, serials, and nonprint materials such as maps, music scores, music, movies, pictures, and other electronic literature. The oldest publications listed in the catalog were issued in 1868 and new holdings are cataloged every year. Here, we consider the name authority file of JAPAN/MARC to be a kind of author list of publications. We analyzed the name authority file of July 5th 2008, which lists 681,924 records of individual names. Of these individual names, we filtered the records to select only Japanese names in Japanese characters, which left us with 572,638 names. Figure 4 shows the full name frequency ordered by rank. The left side is illustrated in normal scale and the right side in log scale. This depicts the power-law form as the same as family name distributions. The frequency of the first rank is 29 and the number of unique full names is 527,567. The number of full names whose frequency is 1 is 499,500. The complement of the full names whose frequency is 1 is the set of full names whose frequency is 2 or more. This number is 73,138. The total ratio is 12.77 %, which we call the name ambiguity rate. The number of observed full names N_{full} is assumed to obey the same power-law formulae as family names because the full name frequency shows a power-law form. It can then be expressed as $N_{\text{full}} = N^{0.9938167}$, and the full name distribution can be expressed as $P_{\text{full}}(k) \sim k^{-\gamma}$, where k is the size of the full name (in other words, the number of individuals who have the same family name and personal name) and the exponent $\gamma \approx 3.081039$. These parameters were calculated using the plots shown in Fig. 4.

The full name distribution in Japan is assumed to be similar to that of JAPAN/MARC. The frequency/rank form of the same family name and personal name of a Japanese can therefore be applied to that of Japanese university or institution

[26] Theoretical paper is written by Newman [27].

[27] http://www.ndl.go.jp/jp/library/data/jm.html.

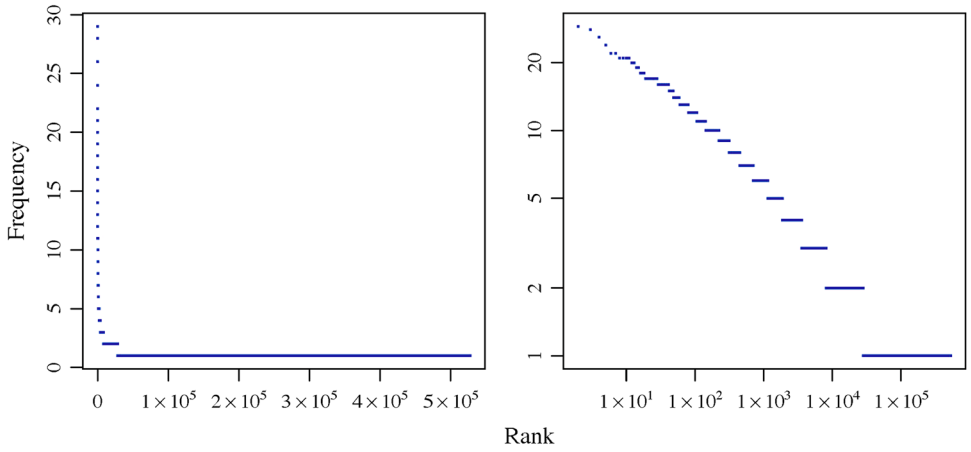

Fig. 4 Full name frequency of JAPAN/MARC name authority file (*left* normal scale, *right* log scale)

researchers. The total number of members in an affiliation corresponds to the sampled population size N of the above formula. When N decreases and gets closer to 1, the proportion of the observed full names gets closer to 1, or simply, $N_{\text{full}}/N \to 1$. When $N \to \infty$, $N_{\text{full}}/N \to 0$. Thus, we find that the smaller the sample population size, the smaller the name ambiguity rate. Once we get N, N_{full}, and its frequency distribution, the name ambiguity rate is calculated by the complement of the full names whose frequency is 1 for all N.

For 86,395 researcher pages of 226 campus directories, we directly matched full names listed on them to the full names and affiliations listed on the RNR researcher pages. When we encountered multiple researcher pages indicated by the same full name in an affiliation, we decided to drop these pages for the candidates to make a relation. As a result, we ultimately created 44,608 links, which represent 51.63 % of the crawled researcher pages. These links are highly precise (a precision rate of nearly 100 %), but the recall rate is low because this method cannot yet completely resolve the name ambiguity problem.

For the error analysis of the links, we manually checked the accuracy of them. We picked a set of links for a campus directory and checked the accuracy of whether each page is for the right person. We chose a campus directory, to which 130 links were made from RNR, and evaluated them by comparing research fields and keywords on campus directory pages and RNR pages. As a result, 130 links were accurate. This evaluation task is essentially difficult because only individual researchers can judge them for their own pages, and the third parties can only detect similarity between pages on academic personality. On recall rate, we easily predict that our method makes it lower because format variations of name and differences of affiliation caused by record update timing impede better recall. However, of 44,608 links, the number of researchers who have the same family name and personal name was 1,697 (3.8 % of the links). Our method resolved these researchers at all.

5 Implementation

We implemented the researcher identifier framework and the Web resource-linking concept as a Web system called the Researcher Name Resolver (RNR). This system consists of the front-end system to search for registered researchers and view their profiles, the Web service for machines to access the researcher profiles, and the backend system to load the researcher profiles. In this section, we describe the researcher page for the front-end system and the Web services of the RNR.

5.1 Researcher page

On the top page of RNR, users can search for researchers registered in the system by filtering using search conditions. In the default search mode, users put keywords in a single field to search for multiple fields of researcher profiles. In the advanced search mode, users specify each keyword in separate multiple fields, e.g., researcher name, RNR identifier, the Grants-in-Aid for Scientific Research researcher number, affiliated institution, department, job title, research fields, and research keywords. The system then returns a list of researchers in Katakana name order as a default search result. Sort order can be changed to English name alphabetical order, Kanji name character code order, and relevance score order in ascendant or descendant. Each snippet leads to the researcher page.

Figure 5 shows a researcher page that describes a researcher profile and provides useful links to external resources related to the researcher. From top to bottom, the page describes the researcher's basic information, direct links, search links, research fields, research keywords, and page URI. The researcher's basic information consists of name, RNR identifier, the Grants-in-Aid for Scientific Research researcher number, affiliated institution, department, job title, and researcher URI. The names can be written in any or all of Kanji, Katakana, and English. The RNR iden-

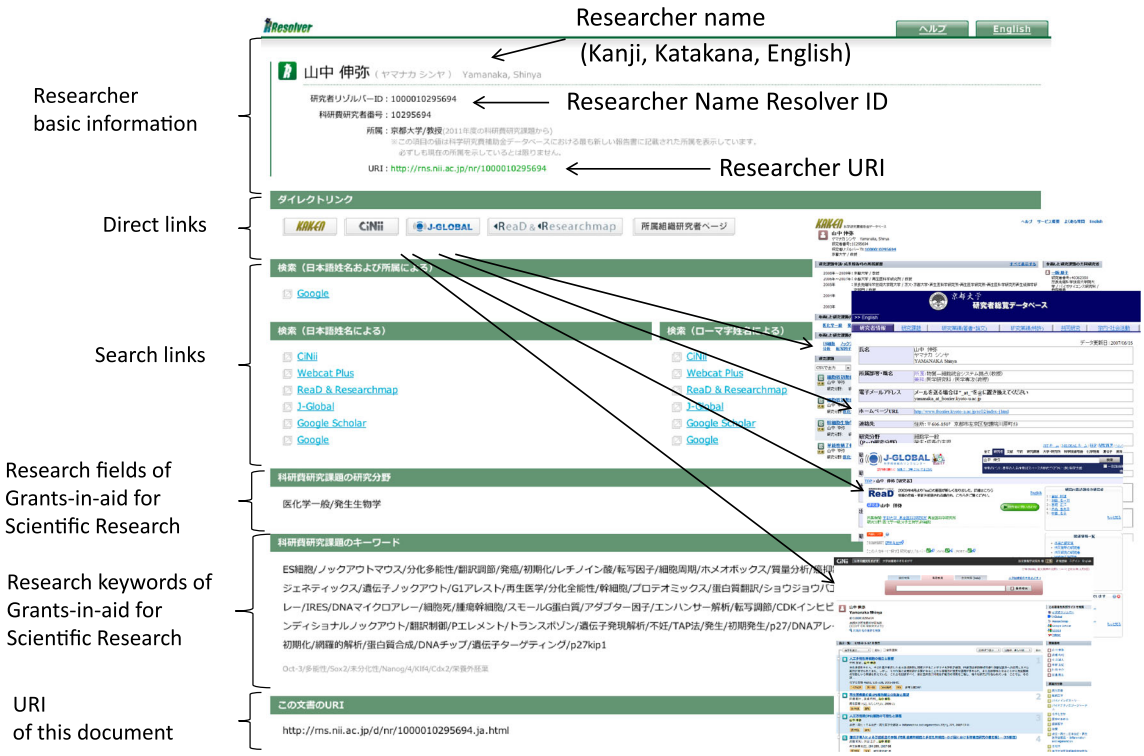

Fig. 5 Researcher page of Researcher Name Resolver

tifier, which is the key in this system, constructs the URI for the researcher that would be referenced by external Web systems.

The most useful element of the page is the direct links, which lead to specific external Web resources related to the researcher. The direct link targets include KAKEN, CiNii Articles, J-GLOBAL, ReaD & Researchmap, and campus directories. The second most useful element is the search links, which lead to the search result of external academic databases whose query format embeds the researcher name and (optionally) affiliation. The result may list irrelevant resources for the researcher, but the one-click action is more efficient than typing a query into Google Scholar. Search link targets include Google, Google Scholar, CiNii Articles, Webcat Plus[28] (an associative search for Japanese books), ReaD & Researchmap, and J-GLOBAL.

Research fields and keywords are necessary parts to determine the academic identity of the researcher. They are extracted from KAKEN researcher profile data.

5.2 Web services

We provide Web-friendly services for external machines to enable easy access to researcher information on RNR. These services are part of a larger technology trend taking place within the online community. We assume that the Semantic Web or Linked Data [24] will be the next-generation data exchange platform on the Web.

To search the registered researchers and the relative documents, we adopted the OpenSearch specification for the API.[29] OpenSearch is a query and response format in XML for search engines. Major search engines provide OpenSearch API. Therefore, most client programmers are familiar with the format.

Researcher Name Resolver currently provides three types of document: an English menu researcher document in HTML, a Japanese menu researcher document in HTML, and a Resource Description Framework (RDF) researcher document in RDF/XML. For the Semantic Web or Linked Data, RDF is a standard description framework to describe relationships among resources. An RDF researcher document describes precisely what a researcher is, as shown in Fig. 6. An oval means a URI resource and a box means a literal resource. Arrows show the relationship between resources. The root of the directional arrow states the subject and the leading edge states the object. The relationship between the two states the predicate.

"http://rns.nii.ac.jp/nr/1000012345678" declared as the "rns:Researcher" type refers to a researcher URI, which is the central concept of this document. The researcher is described

[28] http://webcatplus.nii.ac.jp.

[29] http://www.opensearch.org.

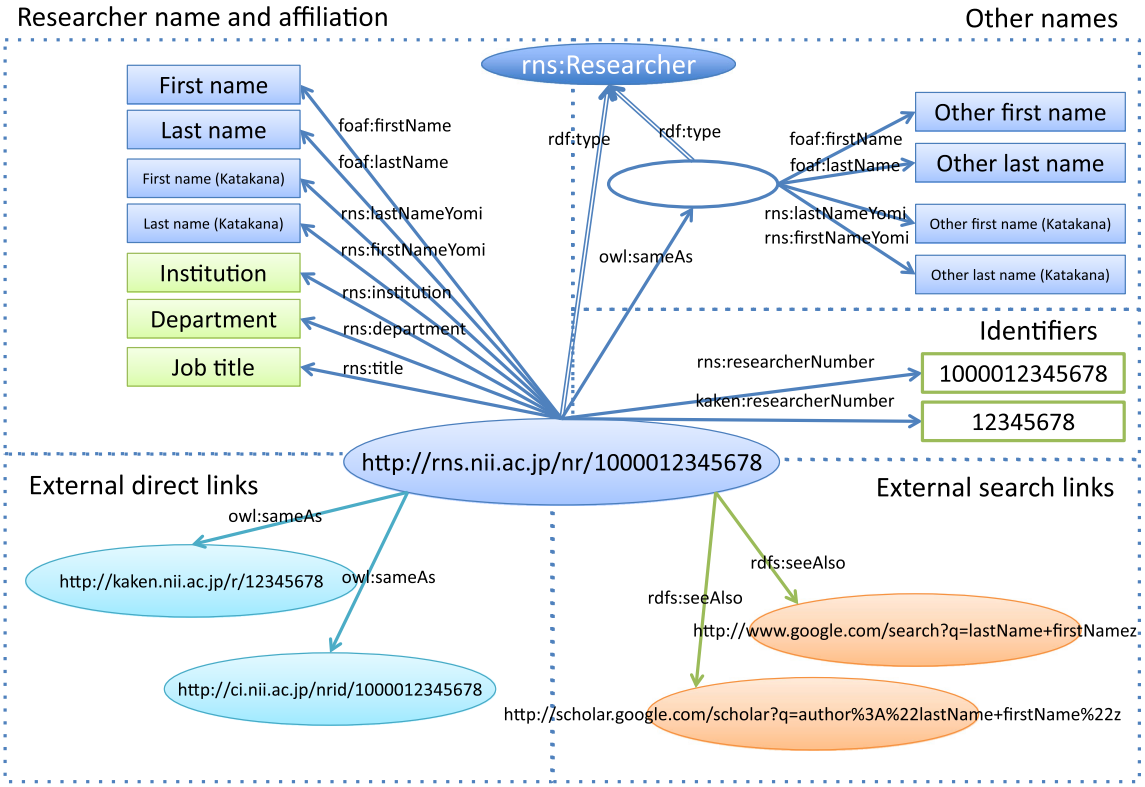

Fig. 6 RDF representation of researcher information

with several attributes, including researcher name and affiliation, other names, identifiers, external direct links, and external search links. Our RDF graph includes the minimum number of attributes required to identify a researcher.

Predicates in the graph consist of a mixture of existing and our defined ontologies. In designing, we tried to adopt existing ontologies first for the researcher attributes, and then, we additionally defined our ontologies for the rest of them. Ontologies are represented by URIs with a form of a prefix, a specific delimiter ":" and a term. An "rns" prefix means that the predicate is defined in RNR while the others mean that the predicate is defined in external sites.

The researcher's name is described in both "foaf:first Name" and "foaf:lastName" for both Japanese and English. Language locale is declared within a literal resource. The researcher name in Katakana is additionally described in "rns:lastNameYomi" and "rns:firstNameYomi" for Japanese phonetic transliterate. Researcher affiliation consists of institution, department, and job title, which are respectively described in "rns:institution", "rns:department", and "rns:title". If a researcher has variations to his or her name that appears in other research papers, these variations are described as alternative names with a "blank node" of "rns:Researcher". This "blank node" is then declared to be the same as the researcher "http://rns.nii.ac.jp/nr/1000012345678" by the "owl:sameAs" predicate.

Researcher identifiers are described in respective predicates. The RNR identifier is described in "rns:researcher Number", and the Grants-in-Aid for Scientific Research researcher number is described in "kaken:researcherNumber". A "kaken" prefix means that the predicate is defined in KAKEN.

The relation between external direct links and external search links and the researcher are respectively denoted by "owl:sameAs" and "rdfs:seeAlso". External direct links construct a network of resources for the same researcher, while external search links are slightly different in that they roughly cover references that might be the relative resources. We therefore adopt a "rdfs:seeAlso" predicate to state the relationship.

For another web service, we provide a URL re-direction service (Fig. 7) based on the fact that RNR manages direct links to external databases. RNR currently contains direct links to KAKEN, CiNii Articles, J-GLOBAL, ReaD & Researchmap, and campus directories. The URL of a direct link usually embeds an identifiable string, i.e., a researcher identifier. RNR maintains these identifiable strings as external identifiers and corresponding URL formats. In our service, a URL specifying a source external database, a researcher identifier of the source external database, and the target external database is re-directed to the researcher page of the target external database. The re-direct destina-

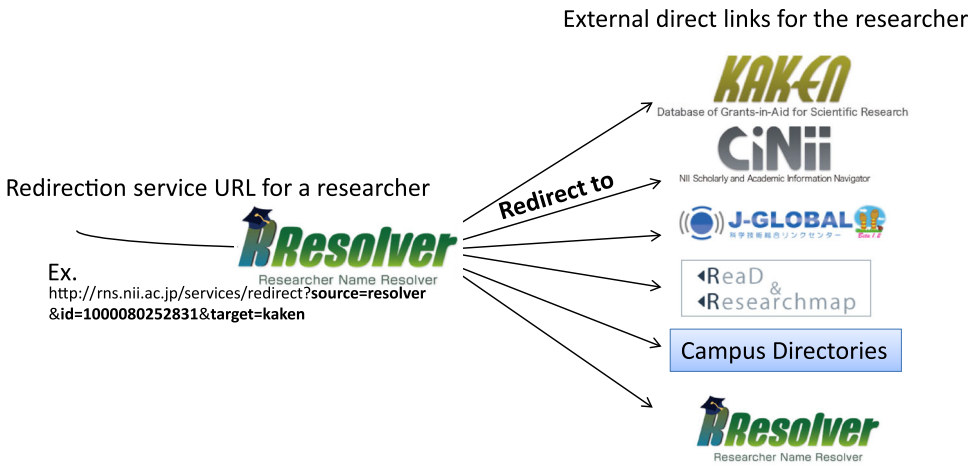

Fig. 7 URL redirection service in which a URL specifying a source external database, a researcher identifier of the source external database, and a target external database is redirected to the researcher page of the target external database

tion can be switched by specifying an external database as the target.

6 Exemplary application for name disambiguation

To determine the feasibility of our concept, we applied RNR to the national search portal of Japanese institutional repositories, namely, the Japanese Institutional Repositories Online (JAIRO). Identifiers are the key to disambiguating author names in JAIRO. We performed a feasibility study in a pilot project with seven university libraries. In this section, we describe how an author name is disambiguated in JAIRO, how an identifier function changes its user interface, and how identifiers are managed in a repository.

6.1 National search portal of Japanese institutional repositories

Institutional repositories tend to be customized for their own purposes. The base systems observed in Japan include DSpace, E-Prints, NALIS-R, iLisSurf, InfoLib, XooNIps, EARMAS, E-repository, WEKO, and T2R2. Of these, DSpace has a 60 % share of the installations. To make author searches possible, some systems include a name authority file that attributes author identifiers valid only in that system. Other systems are connected to a campus directory to exchange author identifiers, profiles, and bibliographic records. Repository managers have gradually recognized the importance of author identifier since DSpace 1.6 implemented an authority control function in 2010. However, there are still a few repositories that provide author searches using identifiers. Author identifier schemes in repositories vary; depending on the scheme, the identifier can be a sequential number, an opaque hash code, an employee identifier, a university member identifier, a national grant researcher identifier, etc. They are applied depending on the legal restrictions and policies of the local organizations.

If we suppose that name authority files for institutional repositories are distributed locally, for the nationwide author search in JAIRO we specifically illustrate the name disambiguation framework in Fig. 8. In this case, RNR functions as the name authority file for JAIRO.

In this framework, repository managers have two options for how they put an author identifier in a bibliographic record. In the first option, they can reference RNR to obtain a URI for a researcher and associate the URI with a creator in the bibliographic record. In the second option, repository managers associate a target researcher URI to a creator in the bibliographic record and upload the researcher profile specifying the URI as an external identifier to RNR. In both cases, a researcher URI is embedded in an "id" attribute of a "creator" field of the junii2[30] metadata, which JAIRO collects via an IRDB[31] harvester. Figure 9 shows example metadata.

If an author identifier would be applicable for a search in a scholarly system, its user interface ought to be changed. We consider JAIRO an example platform for exploring scholarly knowledge, so its user interface ought to be changed along with its exploration behavior. We assume that users adopt one of two potential search modes: the primary mode, in which users use keywords to search for items, or the secondary mode, where users use author names to search for items. These two modes can be switched at any time.

Prototypical user interfaces of JAIRO are shown in Figs. 10 and 11. Figure 10 depicts the top page interface in advanced search mode, which consists of several search fields. One is an author field that will be filled in with an author name or identifier. If a user puts a family name in

[30] http://www.nii.ac.jp/irp/archive/system/junii2.html.

[31] http://irdb.nii.ac.jp.

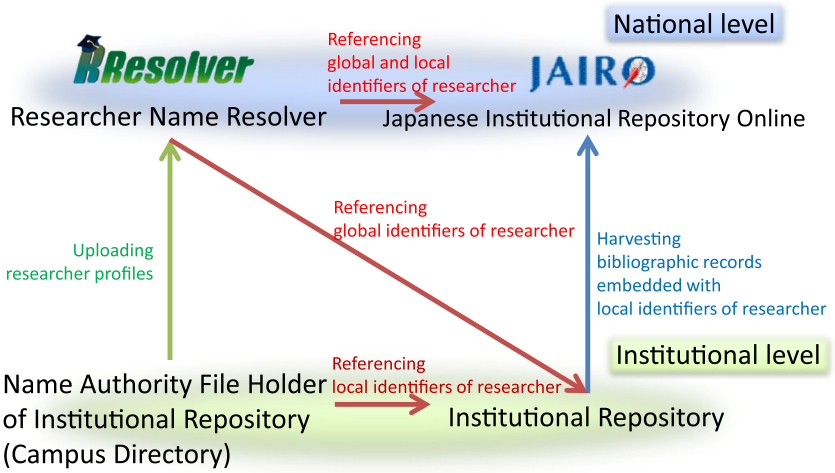

Fig. 8 Name disambiguation framework for JAIRO

```
<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH
  xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2011-05-26T13:34:09Z</responseDate>
  <request metadataPrefix="junii2" verb="GetRecord" identifier="oai:ir.lib.shizuoka.ac.jp:10297/5644">http://ir.lib.shizuoka.ac.jp/dspace-oai/request</request>
  <GetRecord>
    <record>
      <header>
        <identifier>oai:ir.lib.shizuoka.ac.jp:10297/5644</identifier>
        <datestamp>2011-05-22T08:02:22Z</datestamp>
        <setSpec>hdl_10297_24</setSpec>
      </header>
      <metadata>
        <junii2
          xmlns="http://irdb.nii.ac.jp/oai"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://irdb.nii.ac.jp/oai http://irdb.nii.ac.jp/oai/junii2.xsd">
          <title>Orientation-dependent epitaxial growth of GaAs by current-controlled liquid phase epitaxy</title>
          <creator>Mouleeswaran, D.</creator>
          <creator id="http://rns.nii.ac.jp/nr/1000001133354">Koyama, T.</creator>
          <creator id="http://rns.nii.ac.jp/nr/1000048520242">Hayakawa, Yasuhiro</creator>
          <NDC>459</NDC>
          <description>The orientat                of the selective epitaxial growth of Gallium Arsenide (GaAs) has been investigated to achieve a thick epitaxial
layer for application to X-ray                 epitaxial growth was carried out on patterned GaAs with [0 1 1], [0 1 2], [0 1 0], [0 1 -2], [0 1 -1] and their
equivalent seed orientations by current controlled liquid phase epitaxy (CCLPE). SiO2 was used as a mask layer to fabricate the various seed orientations on the
Si-doped GaAs (1 0 0) substrate and various growth periods and                 were considered. Solute transport in the solution was enhanced by the
          <publisher>Elsevier</publisher>
          <NIItype>Journal Article</NIItype>
          <format>application/pdf</format>
          <URI>http://hdl.handle.net/10297/5644</URI>
          <fullTextURL>http://ir.lib.shizuoka.ac.jp/bitstream/10297/5644/1/110520001.pdf</fullTextURL>
          <jtitle>Journal of Crystal Growth</jtitle>
          <issn>00220248</issn>
          <NCID>AA00696341</NCID>
          <volume>321</volume>
          <issue>1</issue>
          <spage>85</spage>
          <epage>90</epage>
          <dateofissued>2011-04-15</dateofissued>
          <language>eng</language>
          <doi>info:doi/10.1016/j.jcrysgro.2011.02.026</doi>
          <rights>Copyright © 2011 Elsevier B.V. All rights reserved.</rights>
          <textversion>author</textversion>
        </junii2>
      </metadata>
    </record>
  </GetRecord>
</OAI-PMH>
```

<creator id="http://rns.nii.ac.jp/nr/1000001133354">Koyama, T.</creator>
<creator id="http://rns.nii.ac.jp/nr/1000048520242">Hayakawa, Yasuhiro</creator>

Fig. 9 A bibliographic record embedded with researcher URI for harvesting

the field, full name variations matching the family name are automatically suggested. An RNR identifier is additionally listed if one is available. A set of a full name and an identifier is the minimum requirement for identifying researchers. In a practical sense, an affiliation should also be included in the record for users to distinguish between researchers.

Figure 11 depicts the list of results for a search specifying an author identifier. In this case, all items in the list belong to the same author, so the users can trace this author's research development path. The grouping function is growing more meaningful. JAIRO provides two axioms for grouping items: item type and institution. Item type grouping provides an understanding of the proportion which type of items an author produces. Institution grouping provides an understanding of the degree to which an author contributes to institutions. If the search condition does not include any author identifier, JAIRO returns a normal result list without using the grouping function.

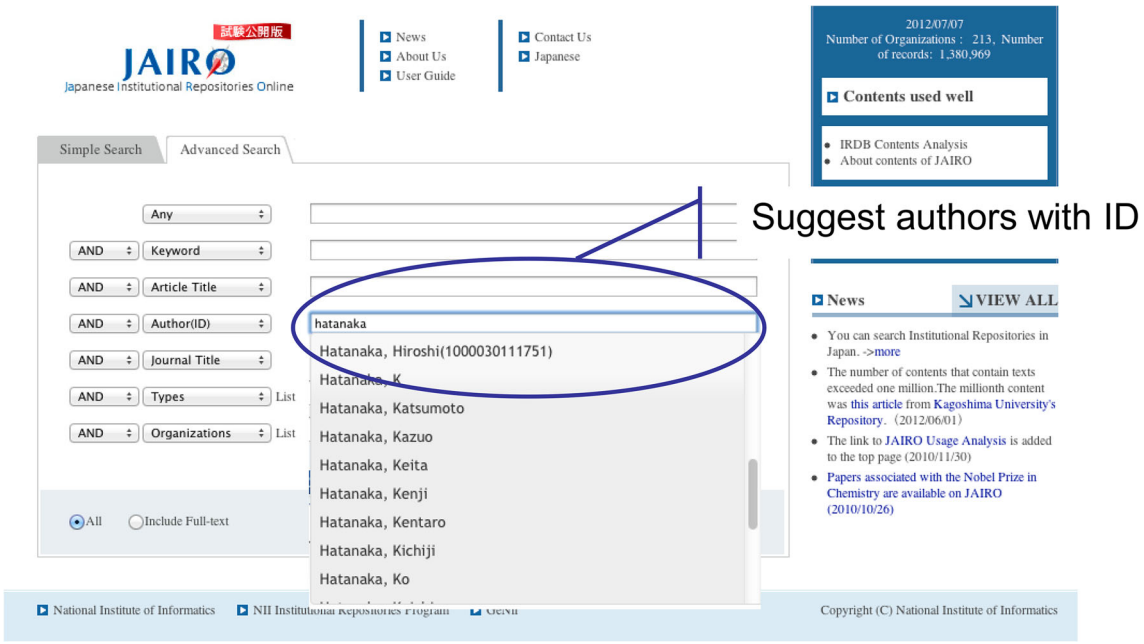

Fig. 10 JAIRO top page in advanced search mode suggesting author names with identifiers

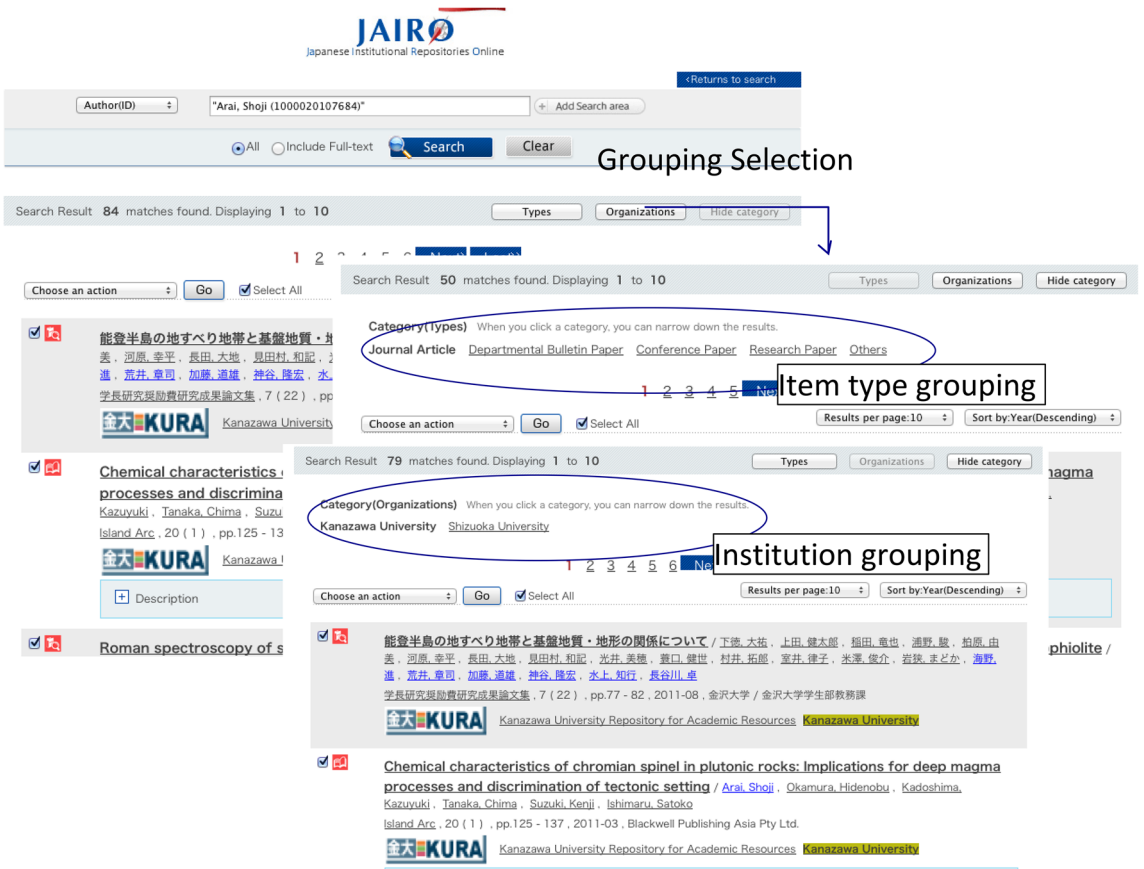

Fig. 11 JAIRO search result grouped by item type or institution

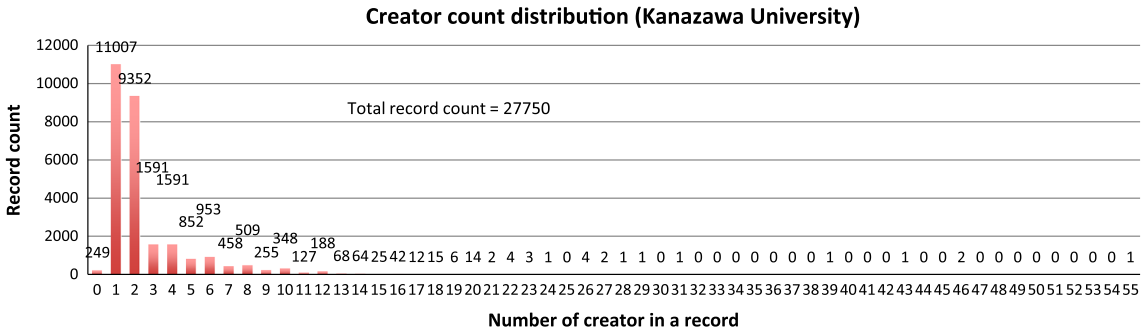

Fig. 12 Creator count distribution in Kanazawa University repository

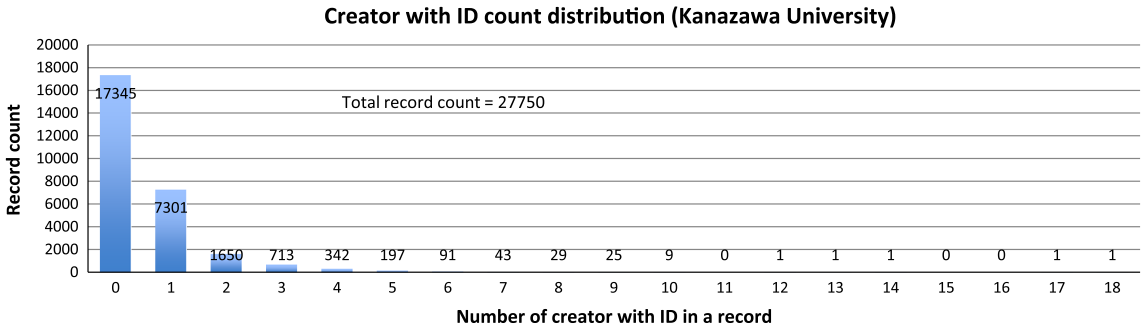

Fig. 13 Creator with identifier count distribution in Kanazawa University repository

We conducted the feasibility study during the 2011 fiscal year using seven university libraries: Kanazawa University Library, Shizuoka University Library, Kansaigakuin University Library, Nara Women's University Library, Nagasaki University Library, Hokkaido University Library, and Osaka City University Library. Of these, six university libraries used DSpace as repository software and one used InfoLib-DBR. Several participants upgraded to DSpace 1.6.0, which has authority control functionality, and prepared a hand-made tool to easily refer to the RNR identifier via the OpenSearch API in the metadata registration workflow. All participants reconfigured crosswalk setting to provide definite bibliographic metadata for this specific task.

Participants assigned author identifiers to creators in bibliographic metadata stored in their repositories to the best of their ability. Kanazawa University Library, which had the highest number of members assigned the Grants-in-Aid for Scientific Research researcher number,[32] uploaded researcher profiles specifying those numbers as external identifiers, and JAIRO collected the bibliographic metadata. Figures 12 and 13 show statistics demonstrating how author identifiers are assigned to creators in the repository. The total number of items recorded was 27,750 and the total number of creators was 71,925. Of these, 16,562 creators were assigned an author identifier, or 23.03 % of the total. Figure 12 shows a creator count distribution per item record. Most item records had either one or two creators, though some exceptional item records had over 20. Figure 13 shows creators with author identifier count distribution per item record. The number of items recorded without any author identifier was 17,345, or 62.50 %. In comparing Figs. 12 and 13, we found that the number of creators with author identifiers was significantly lower than the total number of creators. Ideally, an author identifier is assigned to all creators, but this might not be feasible in actual practice. Librarians might assign author identifiers for the faculty members who belong to the same affiliation and leave blanks for the other authors.

The grand total for all seven university repositories was 106,241 item records and 239,724 creators. Of these, 49,052 creators were assigned author identifiers, or 20.46 % of the total. Nineteen researchers were authors who overlapped two repositories. This fact indicates that researchers contributed their effort to multiple institutions and name disambiguation is required to correctly count their academic contributions.

7 Discussion

In the last section, we discuss for our entire project from the four points of view, i.e., development process, data maintenance, use cases, and future plans of RNR.

[32] http://dspace.lib.kanazawa-u.ac.jp/dspace/.

7.1 System requirements in development

System requirements were not clarified in the beginning of our project. Every time after we released new features of our system, we took some feedback from users and reconsidered required features. The main questions were what kind of researcher profile is shown in public or private, and which a researcher or an administrative staff, who updates the researcher profile, is. We carefully designed the researcher page layout of RNR and decided which researcher profile attributes are appeared on the researcher page. We decided that administrative staffs upload researcher profiles to maintain the authority. We believe that administrative staffs fairly deal with the researcher profiles and used to maintain the researcher profiles in their daily work than researchers. Most of researchers are reluctant to maintain the researcher profiles on such a Web site.

System requirements should be considered from the user needs of a long-time span rather than a short-time span. Researcher identifiers ought to be used for a long time in archival systems, whose essential part of the information architecture suits the needs of a long-time span. These requirements are elicited in a long term of development.

7.2 Data maintenance

The database quality of a researcher identifier management system strongly affects the service quality. Full coverage and error-proneness of identifiers for the target researchers maintain the database quality. Therefore, the quality of identifiers is the important factor to maintain the service quality.

To maintain the highest coverage of identifiers for the target researchers, we re-used the widely used existing researcher identifiers with the highest authority, i.e., the Grants-in-Aid for Scientific Research researcher numbers. The highest coverage of identifiers provided by an authorized organization encourages external systems to reference the identifiers.

In correcting errors on KAKEN database, in which the Grants-in-Aid for Scientific Research researcher numbers are used, researchers report errors and corrections of their identifiers and relative attributes to the authority, and then the data curator in the authority investigates evidences for the corrections based on available documents provided by external authorities. Researcher's requests for error corrections and authority's acceptances for the corrections maintain the truth of the data. After that, RNR reloads the corrected KAKEN-based researcher profile data. This mechanism offers the trustworthy of the service.

For the ResearcherID-based profile data, researchers are not expected to upload their own profile data; rather, this job falls to the organization administrators or to librarians. This ensures the accuracy of the data and saves researchers a lot of boring and tedious work. Japanese academic organizations are already accustomed to collecting researcher profiles and achievement lists and to making staff directories open to the public. They are also cleared to upload batches of researcher profiles and achievement lists to ReaD & Researchmap. In this scenario, administrators or librarians extract researcher profiles and achievement lists from a staff directory, reformat them in our profile format, and upload a batch of the data.

Links for campus directories also need to be maintained because researchers sometimes change their jobs and move to different institutions, and the campus directories are occasionally renewed. The crawler for campus directories needs to be re-activated for catching researchers' movements and re-customized whenever the campus directory structures change.

7.3 Use cases

One of the principal functions of RNR is to provide researcher identifiers on the Web, and its use in combination with external tools is expected to bring out valuable functionalities in scholarly communications. The name disambiguation of authors in the bibliographic metadata of JAIRO is an example of one such functionality.

The Web is essentially a universe with information on everything, so priority concepts for scholarly communications, e.g., bibliographic entities and creator entities, already exist in the universe. Scholarly communication systems such as digital libraries, reference management systems, bibliometrics ranking portals, manuscript tracking systems, and other bibliographic metadata-based systems provide a part of the functionality required for scholarly communications on the Web. Extensively, another type of information systems such as scientific data repositories, software repositories, learning materials repositories, grant databases, and patent databases provides advancements for academic activities on the Web. Data emerge from the distributed sources, are shared among systems, and connections are thus formed between them. Scholarly information essentially requires more qualified connections of information than any other domain.

Researchers, the primary actors of RNR, will cover all researchers in Japan. The identifier scheme is built to have the capability to cover them. The priority actors are the tax-funded researchers in the universities, colleges, and public institutions and national grant recipients in Japan. They will take benefits from the national scholastic systems integrated with RNR while they account their academic achievements based on the systems. Such the systems give the researchers a variety of capabilities, for example, to show clearly their academic contributions to the public, to seek for appropriate research collaborators, and to write curriculum vitae instantly for finding a new job.

The other stakeholders, the actors of RNR, will also get benefits from the integrated systems. University, college, and institution administrators can easily trace academic achievements of their employees. Academic society administrators can effortlessly list candidates of awards with evidence of their contributions. Government administrators of research grant sections can trace research achievements of their grant recipients and utilize the fact for the next policy-making.

7.4 Future plans

Future researcher identifiers may be assigned based on the database maintenance boundaries, which could be local, national, and international. Here, local maintenance corresponds to institutions, national maintenance to government agencies, and international maintenance to international agencies. RNR assigns researcher identifiers on the national level in Japan and provides linkage between institutional and national identifiers. The linkage between international and national identifiers remains to be done. ORCID is the most promising international agency to provide researcher identifiers, so for the next step, RNR should be interlinked with ORCID. This will provide a seamless information surfing functionality via authors of scholarly knowledge on the Web. Users will be able to go through authors integrated with different levels of identifiers and hop to various types of scholarly information services connected to those authors.

International identifiers of ORCID possibly cover national identifiers of RNR, and someday RNR can be replaced with ORCID. After ORCID has released its identifiers in 2012, some UK and US national organizations announced that they would adopt ORCID IDs for national researcher identifiers. However, we expect that this replacement will not immediately happen in Japan at this time of 2013 because national identifiers, i.e., the Grants-in-Aid for Scientific Research researcher numbers, are widely used to identify researchers in the current research administrative systems of Japanese government. Rather, we need linkage between ORCID and the national identifiers, and for further implementations we had better build a bridge function between ORCID and national scholarly systems to exchange researcher profiles.

8 Conclusion

The Researcher Name Resolver (RNR) is an identifier management system designed to cover all researchers in Japan. It is a Web-oriented service to be openly connected with external scholarly systems, and we expect it to be widely used for enriched scholarly communications.

The conceptual framework of RNR consists of a researcher identifier management system and Web resource linking. The researcher identifier scheme is based on the reuse of multiple sets of existing researcher identifiers already listed on the Japanese grant-database KAKEN and the researcher directory ReaD & Researchmap. Researcher identifiers are associated via direct links to related resources on the Web through several methods, including descriptive mapping, focused crawling on campus directories and researcher identification by matching names and affiliations.

We implemented RNR based on this framework. In our implementation, researcher identifiers construct researcher URIs to show researcher pages containing profiles and links to related external resources. Web-friendly technologies, including OpenSearch and the RDF of Linked Data technology, are included to provide Web-friendly services.

We then applied RNR to a name disambiguation task for the search portal of the Japanese Institutional Repositories Online (JAIRO) to determine how the researcher identifier management system cooperated with external systems.

Finally, we discussed the development process, data maintenance, use cases, and future plans of RNR. Through the entire project, we learned important lessons for building an identifier management system. The needs of a long-time span are essentially important and the system requirements are elicited in a long-term development. For the practical operations, data maintenance is required. We found that error-proneness and full coverage of researcher identifiers affect the service quality. Several use cases for the central functionality of RNR showed that it will enhance scholarly communications among researchers. In the future, we will consider international data linkage with ORCID and provide a bridge function for national scholarly systems to exchange profile data with ORCID.

Acknowledgments We are grateful to Tomoaki Sugiyama at Shizuoka University Library for discussions on the metadata framework of JAIRO. We also thank all the developers and support staff who implemented our system.

Appendix: Name ambiguity rate of campus directories

We analyzed the name ambiguity rate of 32 Japanese campus directories. These directories register the university staff who have pages identified by a URI and label them with their full names. The following are statistics from 2009.

University	Registered number N	Unique full names N_F	Number of full names (frequency = 1) $N_F(\text{Freq} = 1)$	Complement of $N_F(\text{Freq} = 1)$ for N	Name ambiguity rate $\{N - N_F(\text{Freq} = 1)\}/\text{N}$
Nagoya Univ.	2,028	2,024	2,020	8	0.0039
The Univ. of Tokyo	2,887	2,883	2,879	8	0.0028
Osaka Univ.	1,042	1,041	1,040	2	0.0019
Kyushu Univ.	2,307	2,306	2,305	2	0.0009
Tohoku Univ.	2,383	2,378	2,373	10	0.0042
Tsukuba Univ.	1,710	1,710	1,710	0	0
Muroran Inst. of Tech.	199	199	199	0	0
Ibaraki Univ.	544	544	544	0	0
Saitama Univ.	420	419	418	2	0.0048
Tokyo Univ. of Agr. and Tech.	506	506	506	0	0
Tokyo Inst. of Tech	1,188	1,188	1,188	0	0
Univ. of Electro-Communications	339	339	339	0	0
Yokohama National Univ.	641	640	639	2	0.0031
Nagoya Inst. of Tech.	367	367	367	0	0
Shiga Univ.	234	234	234	0	0
Osaka Kyoiku Univ.	284	284	284	0	0
Nara Women's Univ.	220	220	220	0	0
Hiroshima Univ.	1,787	1,787	1,787	0	0
Ehime Univ.	864	862	860	4	0.0046
Fukushima Medical Univ.	240	240	240	0	0
Hiroshima City Univ.	190	190	190	0	0
Wayo Women's Univ.	97	97	97	0	0
Keio Univ.	3,207	3,202	3,197	10	0.0031
Kogakuin Univ.	210	210	210	0	0
Sophia Univ.	529	529	529	0	0
Tamagawa Univ.	309	309	309	0	0
Nihon Univ.	3,082	3,073	3,064	18	0.0058
Hosei Univ.	654	654	654	0	0
Meiji Univ.	977	977	977	0	0
Ritsumeikan Univ.	914	913	912	2	0.0021
Osaka Sangyo Univ.	263	263	263	0	0
Ritsumeikan Asia Pacific Univ.	136	136	136	0	0
Max	3,207	3,202	3,197	18	0.0058
Min	97	97	97	0	0
Average	961.19	960.13	959.06	2.13	0.0012

References

1. Enserink, M.: Scientific publishing. Are you ready to become a number? Science **323**, 1662–1664 (2009). doi:10.1126/science.323.5922.1662
2. Hellman, E.: Authors are not people: ORCID and the challenges of name disambiguation. http://go-to-hellman.blogspot.jp/2010/05/authors-are-not-people-orcid-and.html. Accessed 21 May 2012 (2010)
3. Nature: Credit where credit is due. Nature **462**, 825 (2009). doi:10.1038/462825a
4. Bennett, R., Hengel-Dittrich, C., ONeill, E.T., Tillett, B.B.: VIAF (Virtual International Authority File): Linking the Deutsche Nationalbibliothek and Library of Congress Name Authority Files. Int. Cat. Bibliogr. Control **36** (2007)
5. Tillett, B.B.: Authority control: state of the art and new perspectives. Cat. Classif. Q. **38**, 23–41 (2004). doi:10.1300/J104v38n03_04
6. Ginsparg, P.: First steps towards electronic research communication. Comput. Phys. **8**, 390–396 (1994)
7. Warner, S.: Author identifiers in scholarly repositories. J. Digit. Inf. **11**, 10 (2010)
8. Krichel, T.: About NetEc, with special reference to WoPEc. Comput. High Educ. Econ. Rev. **11**, 19–24 (1997)
9. Lynch, C.A.: Institutional repositories: essential infrastructure for scholarship in the digital age. ARL **226**, 1–7 (2003)
10. Day, M.: Prospects for institutional e-print repositories in the United Kingdom (2003). http://eprints-uk.rdn.ac.uk/project/docs/studies/impact/. Accessed 22 May 2012
11. Salo, D.: Name authority control in institutional repositories. Cat. Classif. Q. **47**, 249–261 (2009). doi:10.1080/01639370902737232
12. Hill, A.: What's in a name? Prototyping a name authority service for UK repositories. In: Proc 10th Int Conf Int Soc Knowl Organ ISKO 08, pp. 1–8 (2008)
13. Dijk, E., Baars, C., Hogenaar, A., van Meel, M.: NARCIS: the gateway to Dutch scientific information. ELPUB2006 Conf. Electron. Publ. Bansko, Bulgaria, pp. 49–57 (2006)
14. Mena-Chalco, J.P., Marcondes Cesar Junior, R.: scriptLattes: an open-source knowledge extraction system from the Lattes platform. J. Braz. Comput. Soc. **15**, 31–39 (2009). doi:10.1007/BF03194511
15. Gatenby, J., MacEwan, A.: ISNI: a new system for name identification. Inf. Stand. Q. **23**, 4–9 (2011). doi:10.3789/isqv23n3.2011.02
16. Berners-Lee, T., Bray, T., Connolly, D., et al.: Architecture of the World Wide Web, vol. 1 (2004). http://www.w3.org/TR/webarch/. Accessed 25 May 2012
17. Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific Web resource discovery. Comput. Netw. **31**, 1623–1640 (1999). doi:10.1016/S1389-1286(99)00052-3
18. Avraam, I., Anagnostopoulos, I.: A comparison over focused web crawling strategies. In: IEEE 2011 15th Panhellenic Conf. Informatics, pp. 245–249 (2011)
19. Cafarella, M., Cutting, D.: Building nutch. Queue **2**, 54 (2004). doi:10.1145/988392.988408
20. Baek, S., Kiet, H., Kim, B.: Family name distributions: master equation approach. Phys. Rev. E **76**, 046113 (2007). doi:10.1103/PhysRevE.76.046113
21. Miyazima, S., Lee, Y., Nagamine, T., Miyajima, H.: Power-law distribution of family names in Japanese societies. Phys. A Stat. Mech. Appl. **278**, 282–288 (2000). doi:10.1016/S0378-4371(99)00546-4
22. Zanette, D.H., Manrubia, S.C.: Vertical transmission of culture and the distribution of family names. Phys. A Stat. Mech. Appl. **295**, 1–8 (2001). doi:10.1016/S0378-4371(01)00046-2
23. Kim, B.J., Park, S.M.: Distribution of Korean family names. Phys. A Stat. Mech. Appl. **347**, 683–694 (2005). doi:10.1016/j.physa.2004.08.028
24. Sauermann, L., Cyganiak, R.: Cool URIs for the semantic web (2008). http://www.w3.org/TR/cooluris/. Accessed 2 Jul 2012
25. Fenner, M.: ORCID: unique identifiers for authors and contributors. Inf. Stand. Q. **23**, 10 (2011). doi:10.3789/isqv23n3.2011.03
26. Hill, A.: Report on national approaches to researcher identification systems (2011). http://ie-repository.jisc.ac.uk/567/1/Report_on_approaches_taken_in_national_researcher_name_authority_initiatives.pdf. Accessed 23 May 2012
27. Newman, M.: Power laws, Pareto distributions and Zipf's law. Contemp. Phys. **46**, 323–351 (2005). doi:10.1080/00107510500052444