

CiNii: Bringing Linked Data to Japan's Largest Scholarly Search Engine

Ikki Ohmukai
National Institute of Informatics,
Japan
i2k@nii.ac.jp

Hideaki Takeda
National Institute of Informatics,
Japan
takeda@nii.ac.jp

Keywords: metadata; linked data; academic information services; OpenSearch; RDF.

Abstract

National Institute of Informatics operates "CiNii" (<http://ci.nii.ac.jp/>), the largest scholarly search engine in Japan. CiNii is a database of journals and proceedings. It stores full text and bibliography of over 3 million articles. In cooperation with National Diet Library, electronic journal publishers and institutional repositories, CiNii also collects bibliographic metadata of 20 million articles. The number of page views was over 10 million in December 2008 and still increasing. Figure 1 shows screenshots of CiNii.

CiNii provides search functions for scholarly articles. Search result shows a list of articles corresponding to the query. CiNii also publishes "Bibliography Permalink", which is a web page describing the bibliography of every article stored in the database. It shows not only title and authors, but also the abstract and a list of references and citations.

To become widely accepted by advanced users and developers, we are continuing to enhance CiNii, e.g., introducing permalinks and being the target of major search engines. In April 2009, CiNii redesigned and relaunched, and Linked Data (Berners-Lee, 2006) is offered for public use of scholarly information. In this paper we describe details of our Linked Data.

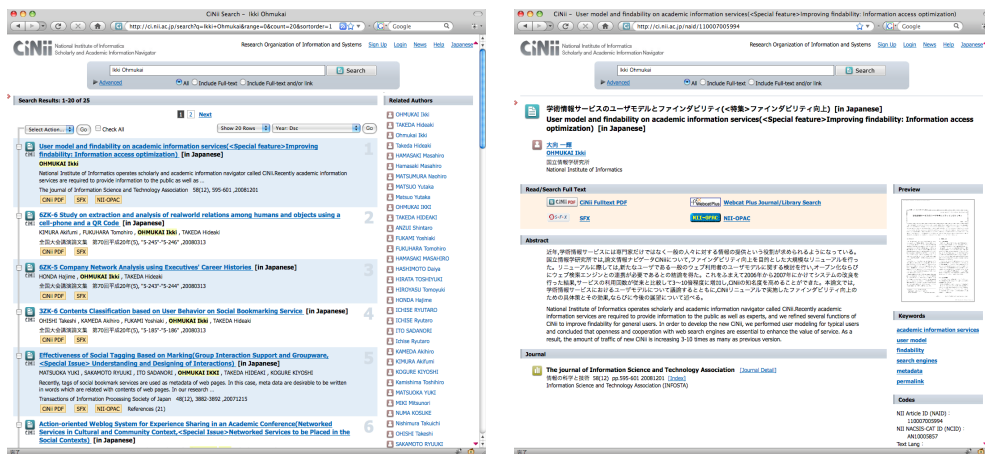


FIG. 1. Screenshots of CiNii.

Summary of Linked Data in CiNii is shown in Figure 2.

Search function in the new CiNii is compatible with the OpenSearch specification. Response formats of OpenSearch consist of XHTML, RSS 1.0 and Atom 1.0. To improve interoperability with other Web services, CiNii does not use original modules, and only uses standard vocabularies such as Dublin Core and PRISM.

In OpenSearch, RSS has a link to bibliography RDF by `rdfs:seeAlso` so that software can obtain machine-readable bibliography without parsing XHTML. Bibliographic metadata is

described with RDF. Like OpenSearch responses, Bibliography RDF does not contain original vocabulary.

There are several problems with representing bibliography by RDF. CiNii cannot know exact URI of full text PDF, which corresponds to the subject of an RDF triple, because the download site might be a different place for different users. So we consider a virtual URI which is combination of bibliography permalink and #article fragment. There is also no reference to URI of RDF itself, but we resolve this problem by introducing foaf:isPrimaryTopicOf to indicate RDF URI.

Author information is expressed by not only dc:creator but also FOAF terms. In FOAF specification, it is impossible to describe that "Person A belongs to Organization B". We use PIM vocabulary (W3C, 2006) to illustrate that statement. There are still some problems. For example, information for more than one person with the same name would be mixed because authors in CiNii do not have unique URI. In the case of affiliation, we define organization name as a resource tentatively because it is not likely for multiple organizations with same name to exist.

We embedded several microformats in XHTML files for quick hacking. xFolk is used in search results, and hCard/hAtom are used in bibliography permalinks.

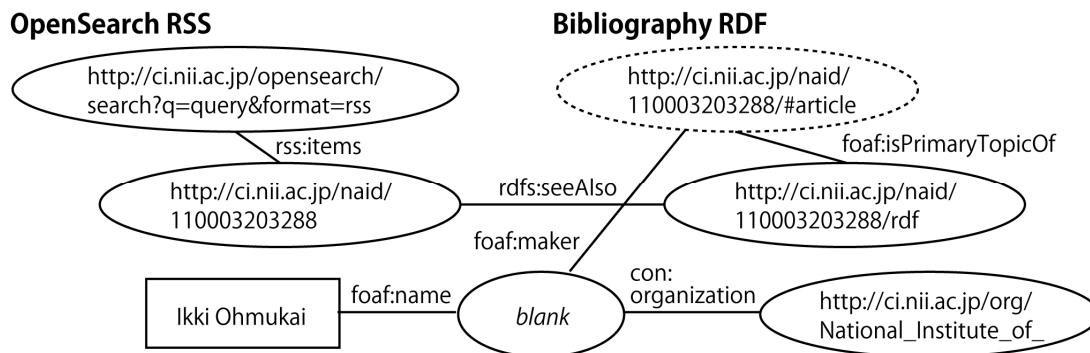


FIG. 2. Model of Linked Data in CiNii.

In this poster we present our efforts to create a large repository of Linked Data from Japan's largest scholarly search engine. We are going to operate the new CiNii and obtain feedback from the users and third-party developers. Trust in our data is still low because of a lack of unique URIs about authors and organizations. We will collaborate with information sources to resolve this problem.

References

- Berners-Lee, Tim. (2006). Linked Data - Design Issues. Retrieved May 7, 2009, from <http://www.w3.org/DesignIssues/LinkedData.html>.
- W3C. (2006). SWAP Personal Information Markup. Retrieved May 7, 2009, from <http://www.w3.org/2000/10/swap/pim/contact#>.