2005 年文学与信息技术国际研讨会 - 论文

关于日本汉文书目网络数据库系统的编程
An implementation of web database system for Japanese *Kanbun* bibliography.

上地 宏一

日本庆应义塾大学
日本二松学舍大学 21 世纪 COE 计划学外合作者
kamichi@sfc.keio.ac.jp

概要 Abstract

In this paper, I introduce a web database system which people can search the bibliography of Japanese *Kanbun* resources by the Internet. The database is produced by Nishogakusha University Japan, a part of 21st Century COE (Centers of Excellence) founded by the Ministry of Education, Culture, Sports, Science and Technology of Japan. There are difficulties for organizing this database, chiefly caused by characteristics of Japanese Language, such as Japanese era name, and several notations (general *Kanji*, traditional *Kanji*, *Hiragana*, and the Latin alphabet), and pen names. The database must allow to include various data of Japanese *Kanbun* bibliography, and allow various search keywords specified by users.

1. 前言

二松学舍大学 21 世纪 COE 计划，以有关日本汉文学的各种各样的文献，论文，杂志的书志检索数据库的构筑作为目标。现在建立检索系统的第一阶段已完成。本论文介绍关于系统的概要和它特有的技术。

2. 什么是日本汉文

所谓本论文的”日本汉文学”，是日本人所记述的汉字汉文的著述等的文献资料作为对象的学问。这个学问包括汉诗等的文学作品，史学文献记录,佛教经典・佛书，天文历法，医书・草本等等领域的文献。

日本人在容纳中国的学术・文化的过程当中，编出了独特的表现形式。比如”训读法”，可以把文法构造和日语不同的古汉语，安上记号和读法，作为日语读。话说在日本汉文学中有名的人物，以江户时代的学者的荻生徂徠和赖山阳为例。赖山阳按照朱子学的思想记下了封建时代的日本历史，名为”日本外史”，在国内外都受高评价。

2.1 二松学舍大学和 21 世纪 COE 计划

二松学舍大学把东洋学，就是日本传统的学问，特别用汉学(中国学)和国学(日本学)的文化做为基本，培养人材。明治当代以后，西洋文化的流入日本，东洋学被轻视。二松学舍大学，进行尊重日本文化的源流，尊重汉文学的教育。

二松学舍大学从 2004 年度开始完成着 21 世纪 COE 计划”日本汉文学的世界的据点的构筑”。21 世纪 COE 计划，是作为为了形成研究教育据点的支援程序作为日本文部科学省的关联团体的日本学术振兴会进行的竞赛的基金。

作为这个”日本汉文学的世界的据点的构筑”计划的一环，总括有关日本汉文学的所有资料的书志数据，以通过互联网能检索书志信息的系统的构筑作为目标。

2.2 日本汉文书志数据库

数据库不仅仅是通常的日本古汉语文献资料，把有关日本汉文学的论文和学会・研究会刊的报道等作为收录的对象。现在是书志数据的输入工作的途中,不过，最终以 100 万件左右的积蓄作为目标。不论日本，世界，作为日本汉文学对象的把研究者作为对象公开这个数据库的检索系统。

3. 为了数据库实现的诸问题

为了收集正在日本全国散逸的日本汉文学资料，构筑数据库，详细决定数据的项目，严密适应的事难。比如，如果有收集由于独自的项目构成的文献目录的形式被保存的书志信息，书志信息可能只有得到信的一部分。这个原故，统一的在在数据项目的决定，年代和书名的数据记述中设置严密的规则事困难。

所以这个数据库，把得到的书志信息直接数据化。并且，在检索积蓄数据的系统方面设法，把必要的数据不缺乏使以能检索书志信息作为目标。

同时，解决了在构筑这个数据库时的一些问题。

[Unicode 和 BUCS]

数据库的字符编码用 Unicode。所以，不但是汉字文化圈的诸语言，日本学研究繁盛的欧洲的学者采用的拉丁系文字也全部都可以数据化，并且能公开于全世界。

为了用 Unicode，也产生新的问题。比如，汉字的异体字处理。日本和中国有不同的文字，在日本国内也有新旧的汉字混在一起的状况。以根据源码分离，字形细小的差异分离的汉字，可能成为弄起检索遗漏的原因。

所以，这个数据库，东京学艺大学的松冈荣志教授制作，用作为日本的工业规格的 JIS 的试行标准的 BUCS，进行异体字的同定。BUCS 关于 Unicode 的汉字把在康熙字典的标题被刊登的字形，或那个最接近的字形看作代表字，列举对代表字的异体字的一览。用 BUCS 能适当解决异体字处理问题。

[由于年号的年的记载]

王朝时代的中国和日本采用年号记述年代。特别日本汉文学的书志信息使用这个年号的信息非常多。可是使用年号的年代记载的检索方法对检索者来说很烦杂。同时，年号的年代记载写法不固定。

所以准备了西历年转换年号记载数据化做的机构。这个机构，转换到西历年各种各样的

年号记载，不仅仅是日本，明代以后的中国和东亚地区的年号记述都可能。

［号、字、别号］

在文献目录的文献资料的著者，有很多本名以外的雅号和字等等，著者名的记载方法没被统一。比如”荻生徂徠”的人物另外有”双松”，”物部”，”物茂卿”等别号。记载这些别号的数据也需要可以检索的状态。

所以，这个系统在内部利用日本国立国会图书馆保有的”著者名根据目录”数据库。这个目录，是在国会图书馆保有的书志信息中，全部收录了同一个人物・团体的别名称的数据库。采用这个数据，有的人名作为检索关键字被输入，准备同样人名的别号的数据，构筑自动检索的机构。

［跟其他机关的数据库的联系］

日本汉文学数据库，把二松学舍大学保有的信息，和全世界存在的所有日本汉文学资料的书志信息的收集作为目的。所以，考虑和被登记于准汉籍的日本汉文学的资料的既存的汉文典籍书志数据库联系。

目前，正在准备能够同时检索京都大学人文科学研究所的”全国汉文典籍数据库”的数据的机构。因为无法区别准汉籍和汉文典籍，”全国汉文典籍数据库”有时候抽出日本古汉语无关的数据，不过，能得到更多的书志信息。

日本汉文学书志数据库，一方公开二松学舍大学独自的检索系统，以他方，国立信息学研究所的 NII-DBR 这个通用学术研究数据库登记。NII-DBR 收录 25 种数据库，140 万件记录。根据这个学术数据库登记日本汉文学的书志数据，日本汉文学研究者以外也期待着更多的研究者触及日本汉文学资料。

4. 系统配置

数据库系统由数据管理部分和，数据公开・检索部分组成，全都是顾客・服务器方式。

数据管理部分，数据库利用 PostgreSQL，管理系统利用着 Java(Tomcat+JSP)。能进行这个原故，数据管理哪里都工作。

关于数据检索部分，现在因为技术的领域的验证阶段，准备着专用的程序。这个程序被 PostgreSQL 容纳的原数据进行各种处理，生成检索用数据文件。这个程序用 Perl 语言制作了。

检索发动机部分用 Perl 语言制作的 CGI。

4.1 检索系统

检索系统，与异体字的同定，著者的别号的同定，使用了西历年的检索对应。

同时，为了汉字的输入困难的欧美研究者，与罗马字输入的检索对应。用罗马字输入的话自动地被日语假名文字转换，进行对数据库内的读音假名项目的检索。

这样，以不仅仅是汉字文化圈的研究者，欧美研究者也能有效的利用的全世界的数据库检索系统的构筑作为目标。

5. 日程和现状的问题点

现在数据库系统的数据管理部分已完成，进行数据登记工作。数据检索・公开部分，满

足了技术方面的检索引擎的第一阶段的构筑。

今后，把以输入中的数据，验证检索系统实用不实用。请更加一部分的对象者，公开系统实际利用。此后，广泛公开。

其次叙述在现状的问题点和今后的预定。

[除去在跟其他机关的联合中不适当的数据]

以前的古汉语文献数据库，常常没明确的区别汉文典籍和准汉文典籍。所以，与其他机关的数据库联系日本汉文学数据库的时候，有与日本汉文学直接没关系的数据搀混了的可能性。关于除去这个无用的数据的方法，现在正在讨论。

[关联书籍(同名，丛书)的表示]

这个数据库不是单纯的书志数据库，以更高度的检索机能的扩充作为目标。对用户指定了的有的文献，考虑表示那个关联的文献的检索机能。

[版面画像，全文文本的追加]

现在收录着只书志信息,不过，最终考虑封面・版本记录等的版面的图像文件的收录，给数据库的全文文本的收录。

6. 结论

这次介绍了在二松学舍大学 21 世纪 COE 计划中关于构筑中的日本汉文学书志数据库。现在还是构筑阶段,不过，为了制作成为听许多人的意见，以全世界规模实用的系统努力。

同时，现在以全世界规模收集着日本汉文学资料的书志数据。诸位的所属机关如果保有有关日本汉文的数据，我们希望无论如何让日本汉文数据库就登记。同时，已经构筑着数据库，协议想跟二松学舍大学的数据库的联合联系的可能性。