		Date : 31/08/2006
		VIAF (Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress Name Authority Files
		Rick Bennett OCLC Online Computer Library Center Dublin, Ohio USA
		Christina Hengel-Dittrich Die Deutsche Bibliothek Frankfurt am Main Germany
		Edward T. O'Neill OCLC Online Computer Library Center Dublin, Ohio USA
		Barbara B. Tillett Library of Congress Washington, D.C. USA
Meeting:	123 Cataloging	
Simultaneous Interpretation:	Yes	
<p>WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL</p> <p>20-24 August 2006, Seoul, Korea</p> <p>http://www.ifla.org/IV/ifla72/index.htm</p>		

Abstract

Die Deutsche Bibliothek, the Library of Congress, and OCLC Online Computer Library Center are jointly developing a virtual international authority file (VIAF) for personal names which links authority records from the world's national bibliographic agencies and will be made freely available on the Web. The goals of the project are to prove the viability of automatically linking authority records from different national authority files and to demonstrate its benefits. The authority and

August 12, 2006

bibliographic files from the Library of Congress and Die Deutsche Bibliothek were used to create the initial VIAF which contains over six million names with over a half million links. A key aspect of the project was the development of automated name matching algorithms which use information from both authority records and the corresponding bibliographic records. The practicality of algorithmically linking the personal names between national authority files was demonstrated; seventy percent of the authority records for personal names common to both files were automatically linked with an error rate of less than one percent. The long-term goal of the VIAF project is to combine the authoritative names from many national libraries and other significant sources into a shared global authority service.

Introduction

Several groups within the International Federation of Library Associations and Institutions (IFLA) Section on Cataloguing recognized the potential of a virtual international authority file (VIAF) [1] where authority records representing the same entity from the world's national bibliographic agencies would be linked and made available on the Internet. Such a VIAF would be a practical expansion of the concept of universal bibliographic control and would build on the work done by each national bibliographic agency. It would permit national or regional variations in authorized form to co-exist, thereby supporting worldwide users' needs for variations in preferred language, script, and spelling.

Current proposals for the future of the Web describe the use of ontologies for making the Web more intelligent for machine and automatic processing. The VIAF could be one of the basic building blocks for a "semantic Web" [2] when combined with other controlled vocabularies and authority files from such sources as abstracting and indexing services, archives, museums, publishers, etc. Libraries now have an opportunity to make a great contribution to this future and should help make this vision a reality. It is important to the development of this shared vision that the VIAF be made freely available to users worldwide.

Other projects have looked into linking personal names in authority files. The LEAF Project [3] (Linking and Exploring Authority Files) proposed to link authority records from many different sources, including libraries, archives, documentation and research centers. These records have various formats, and the details of the type and amount of content varies considerably. The LEAF project proposed automatic linking of the records as they are loaded into the system. Due to the diverse sources of name authority records, they found that the only common information that was available for establishing links was the name, including "see-references," and associated dates. Because the name authority records of the current participants frequently don't include dates, the mismatch error rate for their name authority records is expected to be unacceptably high.

The InterParty Project [4] is an EU-funded demonstration project to create linking authority files among diverse organizations for the primary purpose of supporting digital rights management. The proposed InterParty system would provide a single point of access to the multiple databases involved in the system, so it first provides a centralized search service. As links are manually identified between the names in any of the databases, the individual making the association can enter the link. These links can then be used automatically. Depending on the organizations making the links, the links may be considered sufficiently trustworthy. The assertion of a link by one party does not need to be accepted by other parties involved in the system. The project allows for the possibility of algorithmic matching, but does not specify the techniques or data requirements necessary to support the linking capability.

The VIAF Project

During the 2003 IFLA World Library and Information Congress in Berlin, Die Deutsche Bibliothek (DDB), the Library of Congress (LC), and OCLC Online

Computer Library Center (OCLC) agreed to develop a Virtual International Authority File (VIAF) for personal names [5]. The goals of the VIAF project are to prove the viability of automatically linking authority records from different national authority files and to demonstrate the benefits of a VIAF. The VIAF project will link the name authority files of the Library of Congress and Die Deutsche Nationalbibliothek through a single virtual name authority system. OCLC is developing the software to match personal name authority records between the two authority files. The long-term goal of the VIAF project is to link the authoritative names from many national libraries and other authoritative sources into a shared global authority service for persons, corporate bodies, conferences, places, etc.

The VIAF project consists of five phases:

1. Build "Enhanced Authority" records from both Personennormdatei (PND) and LC Authority Records. This will include identification of the appropriate authority records to include in the enhanced authority records and determination of any special handling needs for the incoming files.
2. Develop matching algorithms, and match PND and LC enhanced authority records to create the initial version of the VIAF. This was an iterative process with Phase 1, as intermediate matching results highlighted additional information that could be extracted and included in the enhanced authority records to improve matching.
3. Build an Open Archive Initiative (OAI) [6] server to provide access to the VIAF.
4. To maintain the VIAF database, additions and changes to both the authority and bibliographic records of all participating agencies are required. This update and maintenance system will be designed around the protocols used by the OAI to request this information for the updates.
5. To access the VIAF records, a user interface will be made available on the open Web. Eventually, the database and interface will support Unicode and multi-language, multi-script capabilities. Direct requests to the database, providing for example an LC version name and requesting the matched PND name as a simple HTML link, can be used to support semantic Web capabilities.

The project initially is focused on demonstrating the feasibility of VIAF by linking the personal names authority records between the Personennormdatei (PND) and the Library of Congress Name Authority File (LCNAF). As of December 31, 2005, the LCNAF file contained 4.2 million authority records for personal names. As of the same date, LC had created and distributed a total of 9.3 million bibliographic records.

As of Fall 2005, the PND file contained 2.6 million authority records for personal names. The PND authority file is used in both DDB bibliographic records and the Bibliotheksverbund Bayern (BVB) bibliographic records. Between the two bibliographic files, there are a total of 15 million bibliographic records associated with PND authority records.

The Name Matching Problem

Initially, the VIAF will function as a German-English and English-German dictionary for personal names. For example, for an American user searching for **J. P. De Valk** (the form of the name established by LC), the name could be automatically 'translated' to **Johannes P. De Valk** (the form established by DDB). As in this case, it is common for different international cataloging agencies to establish the names differently or, conversely, to use the same form of the name to represent different authors. It is possible that **J. P. De Valk** could be established by DDB for a completely different author.

Personal names may take different forms for the same person or have the same form for different people, making it difficult to reliably match names from different authority files. The coverage of the two authority files is significantly different; only a small fraction of the personal names are present in both files. Therefore, information other than the name itself must be used to ensure a reliable match. In authority records for personal names, the person's birth and/or death dates are often present. The combination of birth and death dates is usually sufficient to distinguish people with similar names.

To confirm this difficulty in matching authority records without using supplemental information, a sample of common names from the LC and DDB authority files were extracted. These authority records pairs were then manually reviewed to determine whether they represented the same person. This review found that about 10% of the personal name pairings were for different people. Thus, the match error rate using just the established form of the name would be unacceptably high. Since the name forms are not always identical between the two national authority files, pairing similar, but not identical, names would lead to an even higher error rate. This simple approach also fails to match numerous names that had been established differently.

The Name Matching Solution

Additional matching information is clearly needed to confirm or reject potential personal name matches. For example, consider the following LC authority information for Diane Glynn:

```
100 10 $a Glynn, Diane, $d 1946-
400 10 $a O'Connor, Diane, $d 1946- $w nna
670   $a Country western dancing, 1994: $b CIP t.p. (Diane
      Glynn) pub. info. (an avid country w. dancer & co-
      author          of How to make your man more sensitive)
```

The only directly usable data are the names and the date of birth. Two titles are included in the 670 (Source Data Found) field that might be extracted by machine processing. In practice, only some of the titles can be reliably extracted from these fields.

Bibliographic records are an obvious source for additional information about the person. These bibliographic records can be mined for additional attributes about the

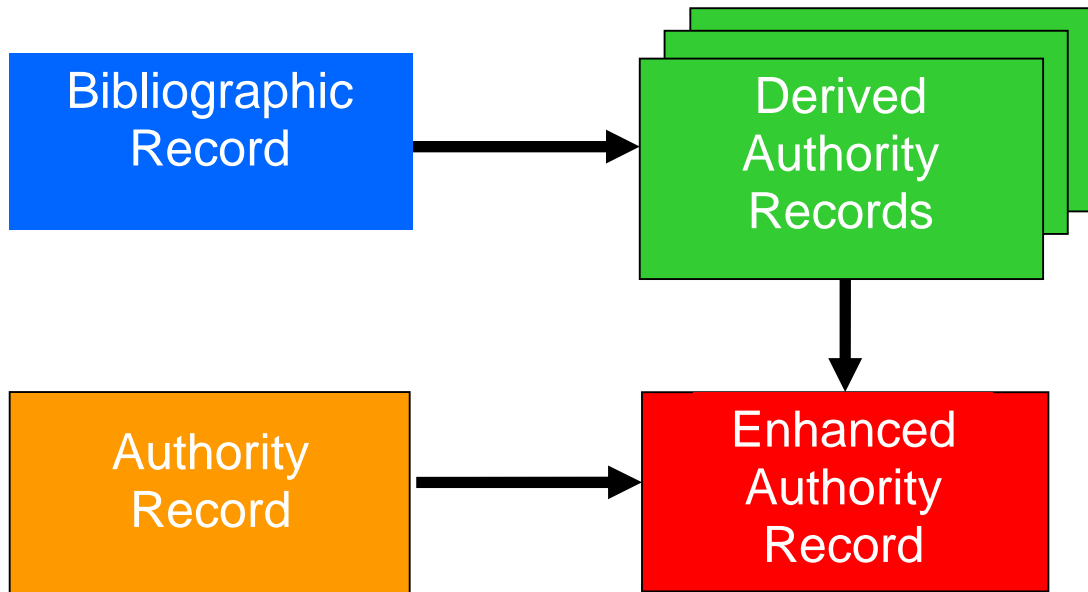


Figure 1. Creating the Enhanced Authority Record

