Виртуальный международный авторитетный файл

Virtual International Authority File[1]

Віртуальний міжнародний авторитетний файл

Барбара Тиллетт
Библиотека Конгресса, Вашингтон, Округ Колумбия, США

Barbara Tillett
Library of Congress, Washington, DC, USA

Барбара Тіллет
Бібліотека Конгресу, Вашингтон, Округ Колумбія, США

Предметом рассмотрения являются инициатива ИФЛА по созданию Виртуального международного авторитетного файла (VIAF) и мероприятия по ее претворению в жизнь.

This paper focuses on the IFLA initiative of a Virtual International Authority File (VIAF) and some activities related to making this a reality.

Предметом розгляду є ініціатива ІФЛА по створенню Віртуального міжнародного авторитетного файлу (VIAF) та заходи, щодо втілення її в життя.

The IFLA Cataloguing Section has been the center of major international standards for cataloguing for nearly 50 years. The most recent standards have been:

- the Paris Principles of 1961 and the current activities to update and expand those principles through the IFLA Meetings of Experts on an International Cataloguing Code (IME ICC),
- the International Standard for Bibliographic Description of 1969,
- the Functional Requirements for Bibliographic Records of 1998, and
- the concepts for a virtual international authority file 2003+.

Background

The concept of a virtual international authority file has been discussed since the 1970's within IFLA — first as part of an idea for a single shared file and more recently in terms of linked national or regional authority files. The virtual international authority file within IFLA is intended to meet several objectives:

- to facilitate the sharing of the workload to reduce cataloguing costs.

Our community has expanded, especially in Europe these days, where libraries are viewed with archives, museums, and rights management agencies as «memory institutions. « We hope authority files could be freely shared among all communities. Shared authority information has the added benefit of reducing the global costs of doing authority work while enabling controlled access and better precision of searching.

Other objectives for authority control are:

- to simplify the creation and maintenance of authority records internationally and
- to enable users to access information in the language, scripts, and form they prefer or that their local library provides for them.

The benefits or virtues of authority control have been debated and restated for decades. When we apply authority control, we are reminded how it brings precision to searches, how the syndetic structure of references enables navigation and provides explanations for variations and inconsistencies, how the controlled forms of names and titles

[1] Also see: "Authority Control on the Web," Barbara B. Tillett. In: Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium : Confronting the Challenges of Networked Resources and the Web, Washington, D.C., November 15-17, 2000. Sponsored by the Library of Congress Cataloging Directorate. Edited by Ann M. Sandberg-Fox. Washington, D.C.: Library of Congress, Cataloging Distribution Service, 2001, p. 207-220.

"A Virtual International Authority File," Barbara B. Tillett. Record of Workshop on Authority Control among Chinese, Korean and Japanese Languages (CJK Authority 3), March 14-18, 2002, held at National Institute of Informatics (NII) in cooperation with National Diet Library. National Institute of Informatics, 2002, p. 117-139 (Also in Japanese, p. 140-153)

"A Virtual International Authority File," presentation by Dr. Barbara B. Tillett for the HKCAN Opening Ceremonies, October 4, 2002. Video: http://hkcan.ln.edu.hk/opening/index.htm

and subjects help collocate works in displays, how we can actually link to the authorized forms of names, titles, and subject that are used in various tools, like directories, biographies, abstracting and indexing services, and so on We can use the linking capability to include library catalogues in the mix of various tools that are available on the Web. Controlling forms used for access and displays provides consistency for users.

There are many technological capabilities that are coming together now and we are really at the brink of making a virtual international authority file a reality.

This is virtual because it is not really a file itself, but a linked system that connects existing Authority Files.

We are also making an historic change to how we view Universal Bibliographic Control (UBC). The IFLA UBC principles for authority control are parallel to those for bibliographic control, namely that:

- each country is responsible for the authorized headings for its own personal and corporate authors and
- the authority records created by each national bibliographic agency would be available to all other countries needing authority records for those same authors.
- Even more, that the same headings would be used worldwide.

In the 1960's and 1970's when this was really catching on, technology had not yet advanced to make such sharing practical on an international level. Plus the lack of funding for an international center to manage such a program prevented that visionary concept from becoming reality. As for the same form being acceptable worldwide, the IFLA developers at that time were primarily from North America and Europe and apparently did not acknowledge the necessity for multiple scripts.

Universal Bibliographic Control — New View

For the past few years a new view of Universal Bibliographic Control is emerging from several working groups within IFLA. This new perspective reinforces the importance of authority control, yet puts the user first It's a practical approach that recognizes a user in China may not want to see the heading for Confucius in a Latinized form, but in their own script. Similarly users in Russian or the Ukraine would want to see the heading in your own script and language.

Yet to still get the benefits of shared authority work and creation of bibliographic records that can be re-used worldwide, we can link authorized forms of names, titles, and even subjects through the authority files of national bibliographic agencies and other regional agencies to create a virtual international authority file. These are several models for how this might work and we need to do more pilot projects of prototypes of thee models to test which would be best to pursue.

In order to be of most use to the library users in each country, the scripts should be the scripts they can read! Names we give to an entity can be expressed in many languages and in many scripts. For example, we could write a Japanese name in English or German with a roman script, in Russian in Cyrillic scripts, or in Japanese (in any of three scripts!) and in many other languages and scripts. Transliteration may serve as a way for some users to be able to decipher records, but much better is the accuracy of using original scripts.

We should now provide at least cross references for variant forms of headings in variant scripts when that is appropriate. We should eventually be able to display the script and form of a heading that the user expects and wants.

I believe that many catalogers within IFLA realize the value of preserving parallel authority records for the same entity. This allows us to reflect the national and cultural needs of our individual users, and at the same time to allow us to set up the syndetic structure of cross references and authorized forms of headings to be used in our own catalogues intended for a specific audience following specific rules. It also allows us to include variants in alternate scripts, at least as cross references for now.

Challenges of International Sharing

In working internationally we quickly see the challenges to sharing authority information. Ideally a particular entity will have the same established form in all authority files. But we know this isn't always true. Different entities may be assigned the same established form of name. Also different forms of the name may be established for the same entity.

When we control all the possible forms variations for the names of an entity and associate them with the bibliographic records for the bibliographic resources that they have some role in creating, producing, or owning, we need to explain that to the user. For example, if the user chooses one form, and we then launch a search, as the Getty Museum and Institute does, that includes all the variant forms and retrieves them all, the user sees records with the variety of names and may find this puzzling. For example, I searched under Lewis Carroll, the author of «Alice's Adventures in Wonderland» in the Getty system. Why am I getting back information about this mathematician, Charles Lutwidge Dodgson, 1832-1898?

You need to tell the user about the variant names used for the same person or corporate body or work.

Alternatively, you could do that by recognizing each «persona» and establishing a name for each one with its variant forms and link them with see also references. So for Charles Dodgson, there would be two authority records — one for himself, the mathematician and one for Lewis Carroll the creator of *Alice in Wonderland.*

As we look at linking we must recognize that different cataloguing rules have differences in what they consider entities — choices are not universal, for example, German rules (*Regeln für die alphabetische Katalogisierung*-RAK) do not recognize that the ships logs can be under an entry for the name of the ship as they can be in AACR2, so the Germans would not have an authority record for ships' names. Similarly for events. For meetings of corporate bodies, the German rules would not create a heading for the entity that AACR2 creates in as a hierarchically subordinate heading for a meeting under the name of the corporate body.

There are also different practices for undifferentiated names — the Germans recently changed their rules to differentiate more names — they more commonly used undifferentiated forms for personal names using just initials for forenames. They still do not require as complete a name or a qualified name to distinguish as the Anglo-American Cataloguing Rules call for. However, even under the same cataloguing rules, say AACR2, when we get more information to differentiate a person, we can make a new authority record to differentiate that person from others groups together under an undifferentiated form of name. This also means that the record for the undifferentiated name can reflect different associated entities over time. And different communities recognize different entities.

Shared Authority Control Initiatives

Over the past few years there have been several projects that help us get closer to providing authority control on a global scale: There have been several projects and initiatives over the past decade in particular. Several projects were sponsored by the European Union, such as the AUTHOR Project that converted a sampling of authority records from the 5 participating national libraries (Belgium, France, Portugal, Spain, and the UK) to the same communication format, UNIMARC. The LEAF project (now coordinated at the Staatsbibliothek zu Berlin) looked at linking authority files for archival purposes using Z39.50 protocols and OAI protocols. The <indecs> and INTERPARTY projects were looking for cooperative work among libraries, museums, archives, and rights management communities in sharing authority information.

Within the International Federation of Library Associations and Institutions, the IFLA MLAR (Minimal Level Authority Records) Working Group identified essential data elements needed in authority records (today we'd call these metadata). This work continues by the IFLA Working Group on FRANAR (Functional Requirements for Authority Numbers and Records and FRSAR for subject authority records.

Within the digital metadata community, there was a Dublin Core «Agents» working group that explored recommendations for dealing with authority information in the digital environment.

Another development over the past few years has been the acceptance of Unicode within the Microsoft tools, such as Windows, that facilitates more global compatibility with multi-scripting.

There is also the initiative of expanding the international cooperative cataloging projects, NACO and SACO, to worldwide users of the *Anglo-American Cataloguing Rules* and *Library of Congress Subject Headings* — also promoting authority control on a global scale.

You are probably aware of several initiatives in Russia to build national authority files. Some identified during the May 2003 International Conference on Authority Control held at the Russian State Library in Moscow, include projects for personal, corporate, and geographic names. The National Library of Russia in St. Petersburg is taking the lead for names of Russian authors, and the Russian State Library is taking the lead for foreign authors whose works have been translated into Russian. For corporate body names, the National Library of Russia will take the lead for names since 1930. For geographic names, the project «RuGeo» is led by the Russian State Library with funds from the Russian Ministry of Culture.

Through such projects we are able to reduce the costs of authority work while improving the potential for more precise searching capabilities for our users.

Authority Control Costs and Automated Systems

The cost of authority work is a concern, and international sharing is just one piece of the solution to help reduce costs. Another is to develop better automated system capabilities to support catalogers doing authority work and to even do some of it automatically. Some local systems already provide us with computer-assisted mechanisms for automatic checking of headings against an existing authority file, and we could see this expanded to start a search against a virtual international authority file, if no match was found locally. Here's where having an international resource will help global costs.

We can also envision the capability of displaying the found matches from a virtual file for a cataloguer to edit or to merge information, if desired, into the local authority record. Some systems now provide community specific retrievals to concentrate on the subject needs of a community in selecting resources for online searches, and other systems like «my library» or «my OPAC» even go beyond that to individualize specific retrievals. Those could build in the authority preferences for user preferred scripts and displays for controlled vocabularies.

We want to have the authorized form preferred by a library as the default offered to most users, but we can also envision offering user-selected preferences through client software, or cookies that let the user specify once what their preferred language, script, or cultural preference is — for example for spelling preferences when cultures have variations, as we do with American and British versions of English.

There are many ways this could actually be applied, and I've suggested several scenarios in earlier papers. Let's quickly take a look at one scenario.

Scenario for Using a Virtual International Authority File

Let's say the cataloger is doing original cataloging and finds there is no authority record in the local file for a heading she needs.

A cataloger types in information in the bibliographic record for a heading. The local system checks the local authority file and finds no match, so it tells the cataloguer that the heading was not found and launches a Web search to the virtual international authority file. Up pops the match with a record created at the National Library of Russia in St. Petersburg. Our cataloguer takes a look and perhaps doesn't want all the information but likes a reference or two and wants a link, so the local system asks the cataloger if she wants the system to create a basic authority record from the one found and to make a link to it…and she clicks on «yes».

The local system would automatically build a local authority record, grabbing the linking information from the virtual authority file — that is the record from St. Petersburg, Russia. The cataloger then adds the MARC field 100, authorized form, according to the locally used cataloging rules, in this case AACR2, and our cataloger can add other fields if needed.

The local system adds a linking 700 field — the MARC format has the 7xx fields in authority records where we can put the linking authorized form and the record control number and the source information for future linking. This linking of authority files would primarily be among the national or regional authority files of national bibliographic agencies — depending on the model we choose. I'll come back to that in a minute.

So we've our cataloger in the United States has added another link in the virtual international authority file to the authorized form following AACR2 — note the record control number for the Library of Congress: (LC) n79072979 — and the Russian record for the same entity following the Russian cataloguing rules in Cyrillic script — note the record control number from the National Library of Russia: (NLR)10326.

Then the local system would update the local bibliographic record using information from the authority record.

When a user comes along, the local system or the «cookies» on the user's system, could specify they want to see the Cyrillic form, and we could display it for them. You can also imagine displaying any script or a Braille keyboard output, or we could provide voice recognition response, built on a user's profile or their «cookie. «

Unicode

Figure 1 is not a VIAF record, but rather is an example of what a Library of Congress authority record might look like with Unicode capability to include original scripts as cross references in a library's catalog. This just gives you an idea of what it might be like. The Library of Congress will be implementing a Unicode-based version of our integrated library system in about a year from now — we hope. There is no particular order to the arrangement of the references, except to place the non-roman scripts following the roman scripts. This example shows English, German, Italian, Chinese, Japanese, Korean, Russian, and transliterations (including Wade-Giles and pinyin for the Chinese, since the Library of Congress just switched to use pinyin).

Tag	I1	I2	Subfield Data
010			ǂa n 80050515
035			ǂa (DLC)n 80050515
040			ǂa DLC ǂc DLC ǂd DLC ǂd NIC
100	0		ǂa Confucius
400	0		ǂa Konfuzius
400	0		ǂa K'ung Fu-tzu
400	0		ǂa Kongzi
400	1		ǂa Kong, Qiu
400	0		ǂa K'ung-tzu
400	1		ǂa K'ung, Ch'iu
400	0		ǂa K¯oshi
400	0		ǂa Konfu´t`si˘i
400	0		ǂa Kongja
400	0		ǂa Kung Fu
400	1		ǂa K'ung, Fu-tzu
400	0		ǂa Confucio
400	0		ǂa Конфуций
400	0		ǂa 孔夫子
400	0		ǂa 孔子
400	0		ǂa 孔丘
400	0		ǂa こうし
400	0		ǂa コウシ
400	0		ǂa 공자
670			ǂa Jakobs, P. M. Kritik an Lin Piao und Konfuzius, c1983: ǂb t.p. (Konfuzius)
670			ǂa Konfu´t`si˘i, 1993: ǂb t.p. verso (551-479 B.C.)
670			ǂa His Gespr¨ache (Lun y¨u), 1910: ǂb t.p. (Kungfutse)
670			ǂa Web connection ǂu http://www.friesian.com/confuci.htm
700	0		ǂa 孔夫子 ǂ5 Natl. Lib. of China

Figure 1. Unicode-enabled Future Authority Record for Confucius

VIAF Models

There are several models for a virtual international authority file.

For a distributed model, a searcher would use a standard protocol like the next generation of Z39.50, or SRU/SRW, to search the independent authority files of participating National Bibliographic Agencies or regional authorities.

Another model is to have one central authority file and link all others to that, so that work would not need to be done by each national bibliographic agency with all other participants in this international universe. A cataloger would then get access to all the authority records for that entity worldwide by a single search of the central file. If there was not match in that central file, a search could then be made with Z39.50 to the other files. The problem with this is what library could take the responsibility and expense of being the central file?

Another model is to have a centralized system but keep the authority files where they are. We may find that this model is the best approach in terms of record maintenance. When we bring in the use of the Open Archives Initiative (OAI) protocol, such a system could use a central server to harvest metadata from the national authority files on their own servers. That information would be refreshed in the central server whenever there are changes in the national files. This means the day to day record maintenance activities continue to be managed as they are now by the National Bibliographic Agency (or regional authority).

I am sure you can imagine other variations of these models. And we need to try them out to see which will be best for us in today's Internet environment.

VIAF Proof of Concept Project

A project is now underway to test this centralized virtual authority file model. In August 2003 the Deutsche Bibliothek, the Library of Congress, and OCLC signed a memorandum of understanding to jointly engage in a

research project to test this VIAF concept — a proof of concept project. It is hoped if this proves successful, it can be the basis for a true Virtual International Authority File that links the world's national and regional authority files as a freely available shared resource.

The first stage of this project links our existing authority records for personal names: LCNAF and the DDB's Personal Name Authority File (PND). It involves testing OCLC's matching algorithms to see how much a computer can do for us and how much will require human intervention for matching and checking. OCLC uses information in the bibliographic records (such as dates of publication, subjects, languages, countries of publication, etc.) and information in authority records (authorized and variant forms of names) to do this matching.

This first stage is nearly completed and the results are being compiled for a report by the end of this year. OCLC received the authority records and the bibliographic records from both the Library of Congress and the Deutsche Bibliothek and has done a first pass to compare and match the records.

The total number of authority records that were in the respective files at the end of 2003 (when the set of records to be matched was made available) is nearly 4 million records at the Library of Congress (LC) and nearly 2½ million records at Die Deutsche Bibliothek (DDB) (which includes personal name records from the Staatsbibliothek in Münich).

During the matching process, both the authority records and bibliographic records were examined. For every personal name used either as a main entry or added entry in the bibliographic records, a *derived* authority record was created. In addition to the name, the *derived* authority record includes a coded summary of the material published based on information from the associated bibliographic records. A bibliographic record with multiple personal names will generate multiple *derived* authority records.

All of the *derived* authority records for a particular person will be clustered with the authority record for the individual.

The contents of the all the *derived* authority records for the individual are added to form the ***enhanced*** authority record. OCLC will use the ***enhanced*** authority records as the database for the VIAF.

In Stage 2 of the project, we are building a searchable database. OCLC is using the open source software, Site Search, to provide access to this database as this part of the test.

As we make the links we may be building one or more servers with this «metadata» — one will be housed at OCLC, probably another at the OCLC European office (PICA) and another at the DDB. We were not planning on having one at LC for this project.

In Stage 3, we want to test the concept of using the Open Archive Initiative (OAI) protocols to do the ongoing maintenance of updating the information in the server by harvesting metadata for new and updated or deleted information in the home authority files.

A possible last stage (Stage 4) to test in the proof of concept would be the end user display capabilities, to switch the user's preferred form of language and script that would be displayed on his or her machine. At this time we believe this stage may be omitted from the project, as the technology is not quite yet ready.

We prefer this model for now, as it seems to hold the most promise for maintenance, but there are concerns about scalability, given the overhead of matching with bibliographic records. However, if it proves successful, we would like to include connections to all the major authority files worldwide, including those from other communities like abstracting and indexing services, archives, museums, publishers, etc. We especially look forward to testing this strategy using non-Latin script authority information. We also want to test it on other types of authority records: such as for corporate names, geographic names, and uniform titles. Subjects might seem appropriate, but we know that abstract concepts often don't translate uniquely from language to language. Therefore, it is unlikely that VIAF model would be extended to include subjects.

We really hope we can preserve local forms of names this way and link different records that use varying cataloguing codes and yet still meet users needs.

For the future, we can envision a shared international authority file being an integral part of a future «Semantic Web. « You may have heard about this in a Scientific American article a few years ago now by Tim Berners-Lee, founder of the Internet. The idea is to make the Internet more intelligent for machine navigation rather than human navigation of the Web. It involves creating an infrastructure of linked resources and the use of controlled vocabularies, they are calling «ontologies. « These ontologies could be used to enable displays in the user's own language and script.

Here's where libraries have an opportunity to contribute to the infrastructure of the future Web — we already have controlled vocabularies in our various authority files. Those would be linked with other controlled vocabularies of abstracting and indexing services, of biographical dictionaries, of telephone directories, and many other reference tools and resources to help users navigate and to improve the precision of searches, so users could find what they're looking for.

You can see that we would also build in the search engines and future tools that as a collective resource would connect us to the entire digital world. All of this, of course, would have built-in, appropriate security and privacy

assurances and ways to identify and acknowledge resources that we can trust and rely on, and somehow, miraculously, all the copyright issues will be resolved, so this is definitely in the future! But it's great to think about the possibilities and opportunities for testing this out and to think about how we can improve upon our dreams.

The Web has brought us a new way to convey information. The new twist is that our catalog — that is our PC where the online catalog is displayed, is also the device for viewing the actual digital objects and connecting to the entire digital world.

A freely available Virtual International Authority File could be used by Web systems to improve the precision of users searches and help enable display of the language and script for names and subjects that a user prefers.