A Virtual International Authority File

Presentation by

Barbara B. Tillett, Ph.D.

Chair, IFLA Division of Bibliographic Control

and

Chief, Cataloging Policy & Support Office, Library of Congress

For the IME ICC, Buenos Aires

17 August 2004

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. http://www.lcweb.loc.gov/catdir/bibcontr/tillett.html

Objectives

- Facilitate sharing to reduce cataloguing costs to libraries, museums, archives, rights management agencies, etc.
- Simplify creation and maintenance of authority records internationally
- Enable users to access information in the language, script, form they prefer

It has often been observed that the current Web is chaotic for finding information. It needs help and we can provide it!

Introducing an element of authority control to the Web environment would help meet these objectives:

-facilitating 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." Is this also true in Asian countries? We hope authority files could be 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...

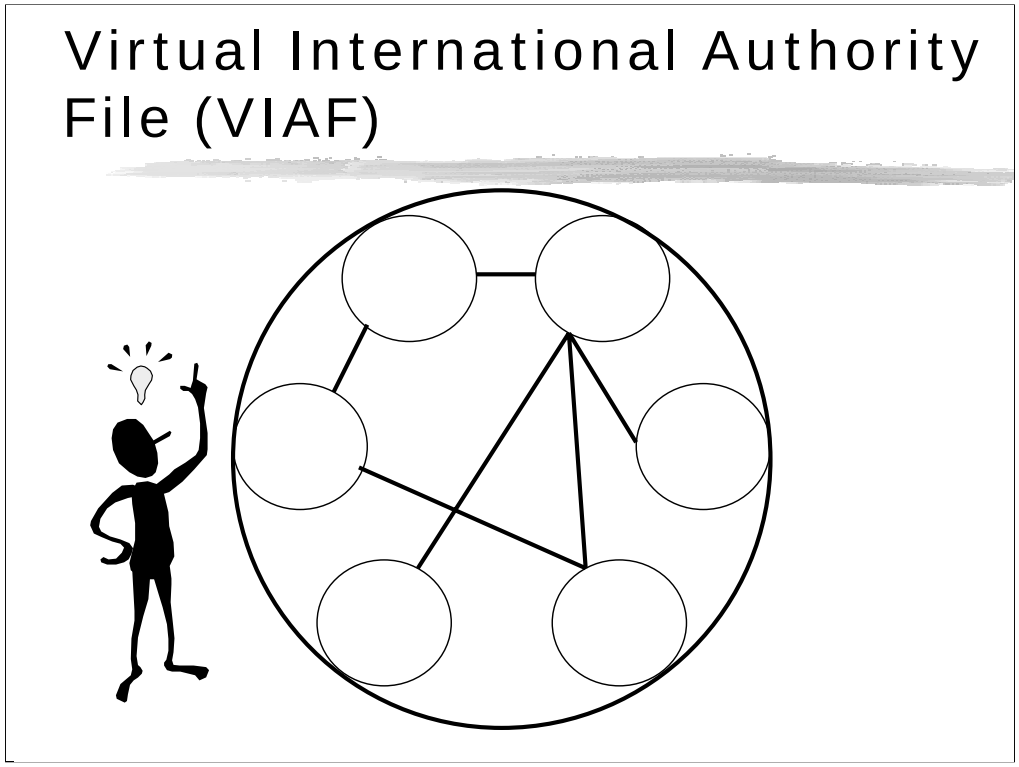
Authority control virtues

- "Precision" in searching
- Syndetic structure of references to help navigate (the variant forms of name/title/subject/etc.)
- Displays to collocate works
- Links to forms used in particular resources
- Bring library catalogues into the mix of tools available on the Web

The virtues of authority control have been debated and restated for decades. When we apply authority control in the Web environment, 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 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.

We are all aware of very poor OPACs that lack cross references or links to authority files and without these features, quite frankly, they are not Catalogues!



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.

IFLA UBC authority principles

- Each country responsible for authority headings for its own personal and corporate authors
- National authority records available for everyone to use
- Same form and structure used worldwide

We're 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.

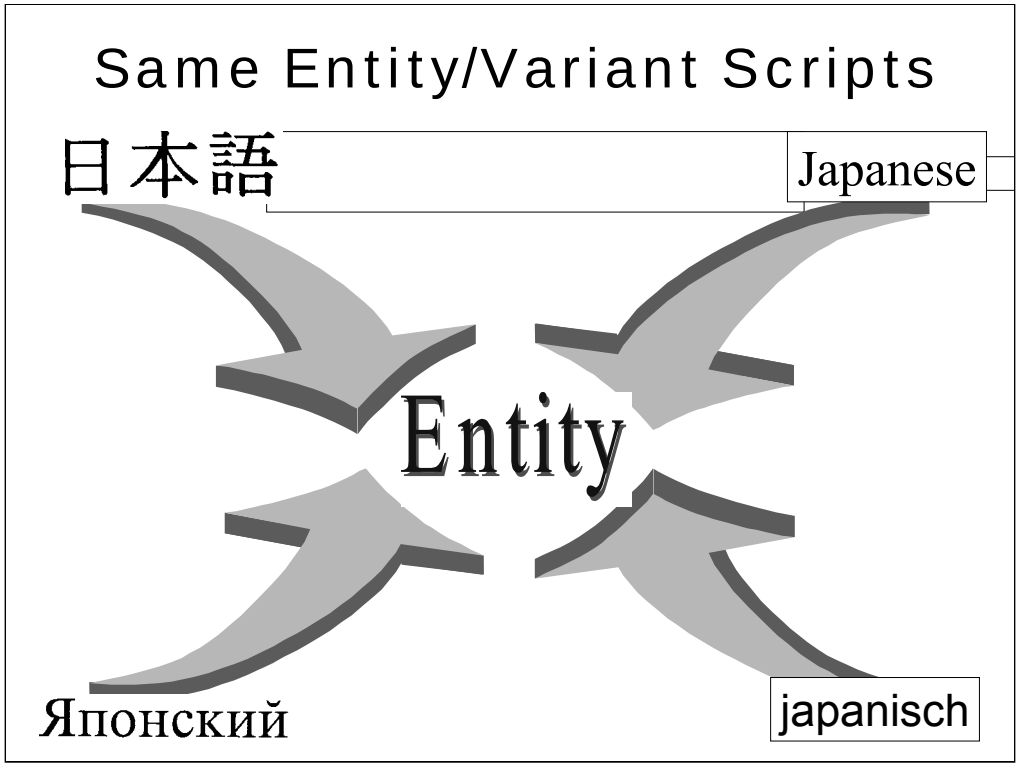
In the 1960's and 1970's when this was really catching on, technology had not yet advanced to make such sharing practical on an international level. Plus the lack of funding for an international center to manage such a program prevented that visionary concept from becoming reality. As for the same form 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.

New view of UBC

- User perspective - to display script/language(s) of one's own country
- Bibliographic agencies still responsible for control in their own countries (or region/cataloguing rules/etc.)
- Link forms established in "national" or "regional" authority files to create a virtual international authority file

For the past couple of 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 Japan or Korea would want to see the heading in their 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! What a novel idea!

This slide shows that the names we give to an entity can be expressed in many languages and in many scripts. For example, we could write it 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. In the United States the group within the American Library Association responsible for changes to the MARC 21 format, known as MARBI, is starting to explore this possibility. More work needs to be done. 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 catalogues intended for a specific audience. It also allows us to include variants in alternate scripts, at least as cross references for now.

Entities

- **Crosswalks and Mappings - not always 1-1**
  - **Examples of differences among cataloguing rules**
    - **Ships**
    - **Events**
    - **Meetings of corporate bodies**
    - **Undifferentiated names**

As we look at linking we must recognize that different cataloguing rules have differences in what they consider entities - AACR2's 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, so they 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.

Programs to facilitate future authority work

- Automatic check of heading against existing local authority file
- If not found, automatic check against “virtual” international authority file
- Display found matches for editing or reference
- Insert authorized forms into local authority record for future link

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 then launch a search against a virtual international authority file, if no match was found locally.

We can also envision the capability of displaying the found matches from the virtual file for a cataloguer to edit or to merge information, if desired, into the local authority record, including capturing the information for future linking.

Switching for displays

- Library default
- User-selected preferences
  - Client set-up, “cookies,” or future method
    - Language
    - Script
    - Culture (country)/spelling
      - Labor vs. Labour

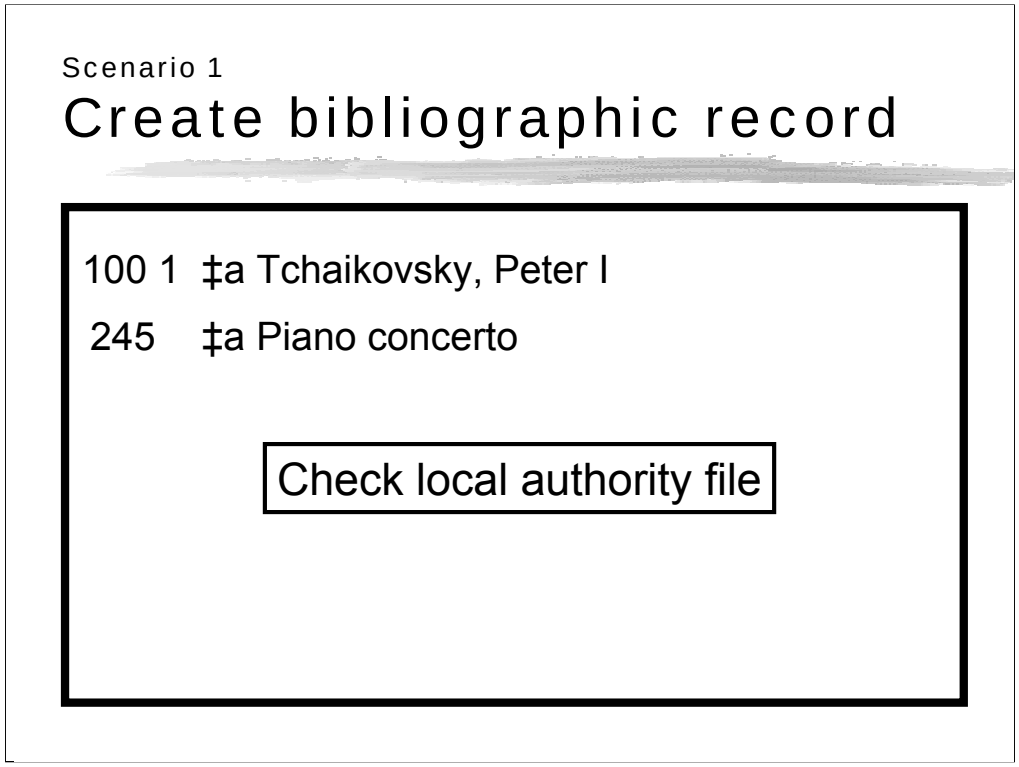
Some systems now provide community specific retrievals to concentrate on the subject needs of a community in selecting resources for online searches, and other systems like “my library” or “my opac” even go beyond that to individual specific retrievals. Those could build in the authority preferences 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, like American English and spelling preferences in the United Kingdom - labor and labour...

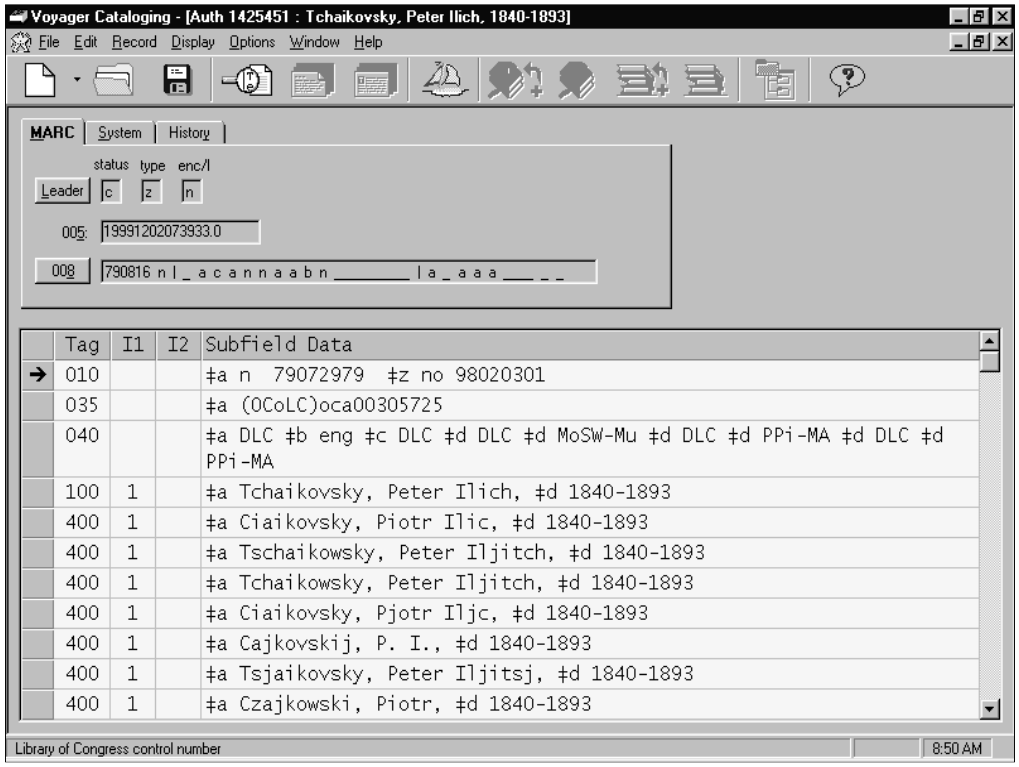
Scenario 1 – Found locally

- Original cataloguing with match found for same entity in local authority file

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 two...



A cataloger types in information into a bibliographic record and the local system checks the local authority file.



The local system found the record in the local authority file and displays it so the cataloguer can confirm it's the same entity.

Scenario 1

Corrects bibliographic record

100 1 ‡a Tchaikovsky, Peter Ilich, ‡ d 1840-1893

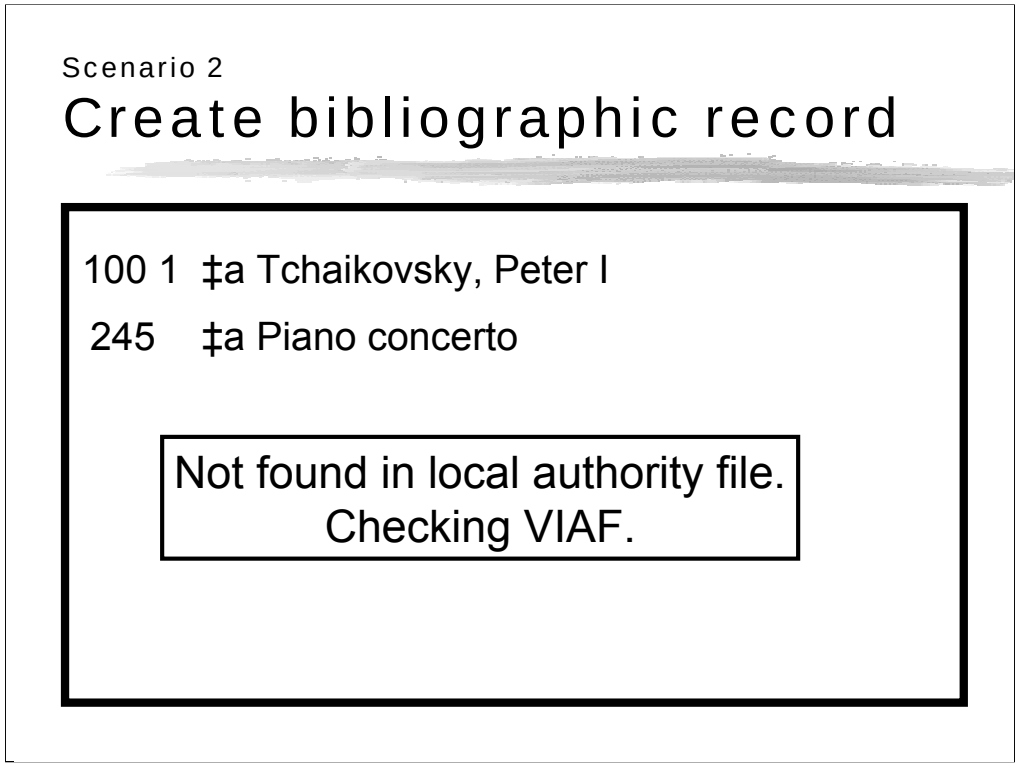
245 ‡a Piano concerto

And we'd like the system then to automatically update the bibliographic record with the authorized information from that authority record, once it is confirmed.

Scenario 2 - Web launch

- Original cataloguing with no match found in local authority file
- Launches Web search of virtual international authority file
  - Assumes response times/system reliability will improve in future

Now what about no record in the local file? Let's look at a second scenario.



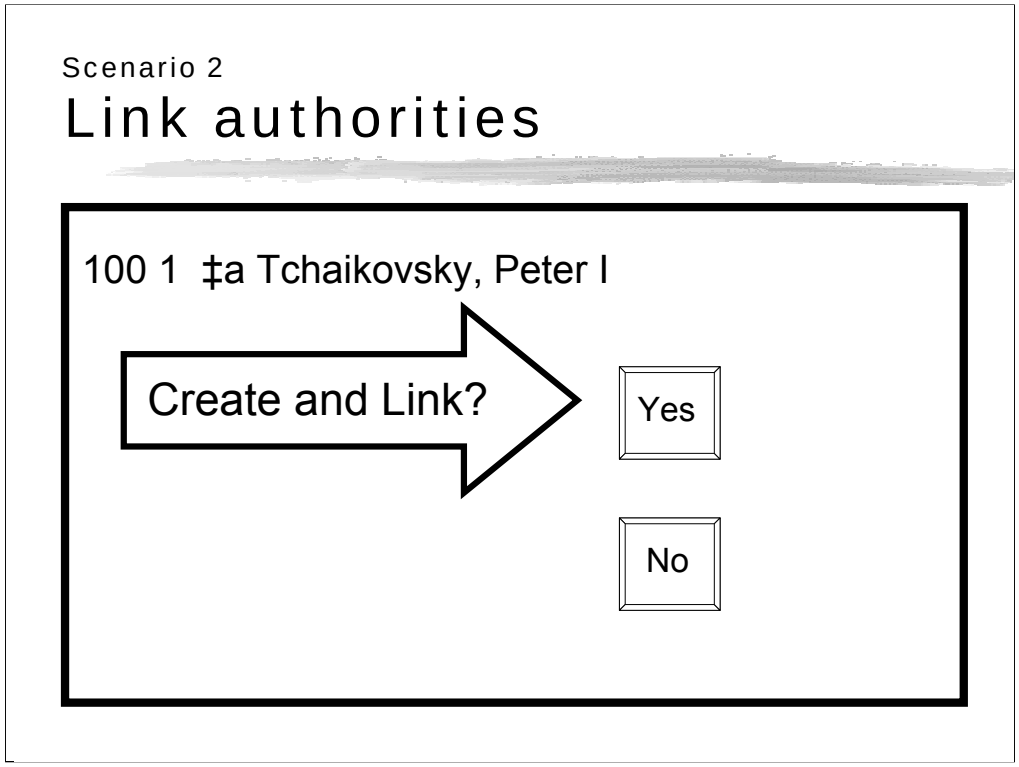
A cataloger types in information. The local system checks the local authority file and finds no match, so it tells the cataloguer that the heading was not found and launches a Web search to the virtual international authority file.

Your search of the Virtual International Authority File found the following match:

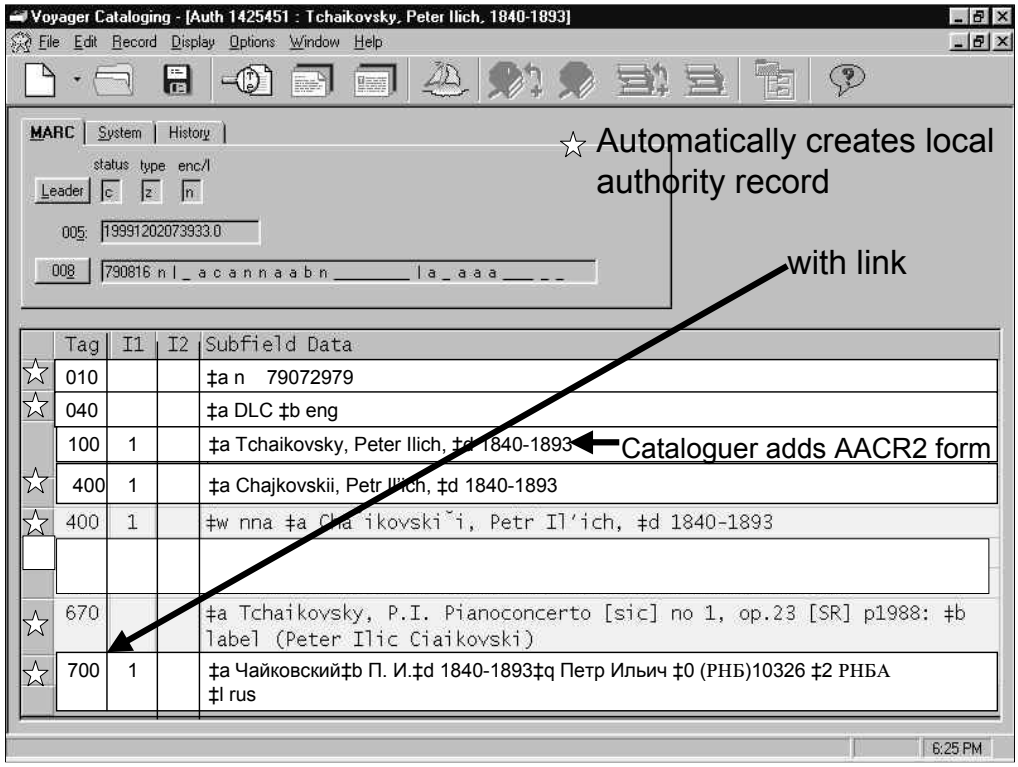
RUSMARC-record
Маркер: 00445nx 22001453 450
001: 10326
005: 20001108144619.0
100: $a20001108arusy0103 ca
200: 1$8rus$7ca$aЧайковский$bП. И.$f1840- 1893$gПетр Ильич$4070
200: 1$8rus$7ba$aChajkovskij$bP. I.$f1840-1893$gPetr Il'ich$4070
300: 0 $aРусский композитор, ученик А.Г. Рубинштейна.
700: 1$310327$8rus$7ba$4070
810: $aВсемирный биографический энциклопедический словарь. - М., 1998.
801: 0$arus$brnb$c20001108
810: $aГАК РНБ.

Up pops the match with a record created at the National Library of Russia in St. Petersburg…. [Aside: The remarkable thing is I did indeed find this record using the Internet and was able to display the roman and Cyrillic characters on my PC!]

Our cataloguer takes a look and perhaps doesn't want all the information but likes a reference or two and wants a link, so

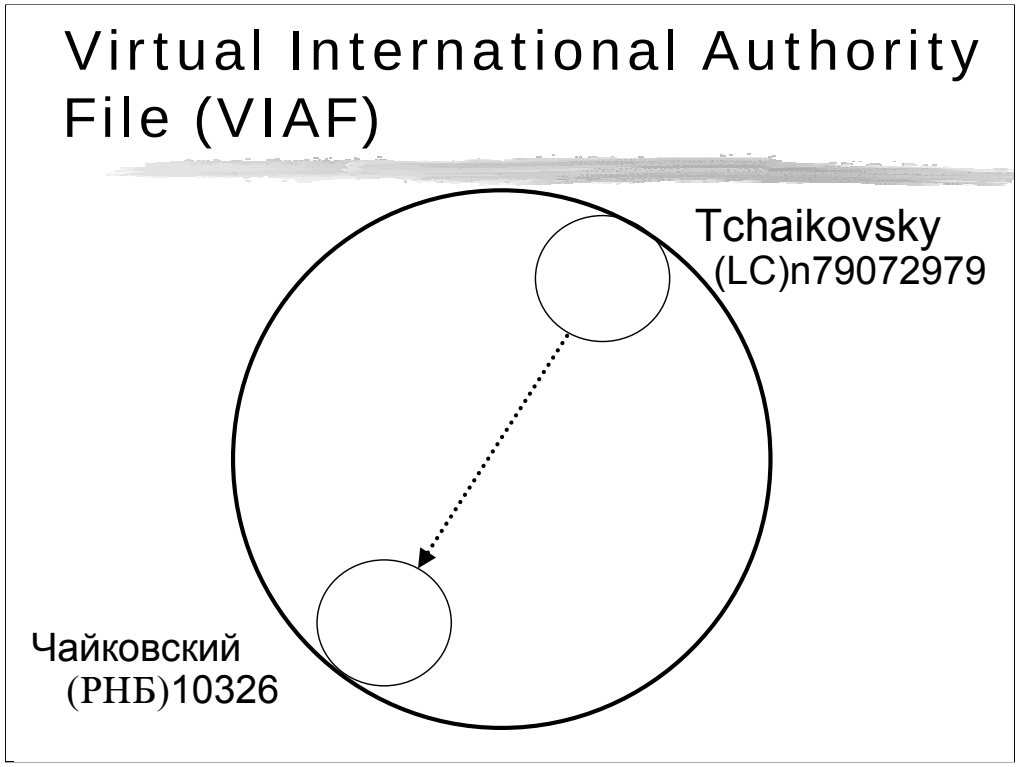


The local system asks the cataloger if she wants the system to create a basic authority record from the one found and to make a link to it…and we click on “yes”.



And our local system automatically builds 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 the 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 now 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: (РНБ)10326.

Scenario 2

Corrects bibliographic record

100 1 ‡a Tchaikovsky, Peter Ilich, ‡ d 1840-1893

245 ‡a Piano concerto

Then our local system updates our local bibliographic record.

Local system - User view

User's cookie specifies Cyrillic/Russian preferred.

Display 700 form, building on local system's authority structure

Чайковский, Петр Ильич, 1840-1893

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 (click) and (click) 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."

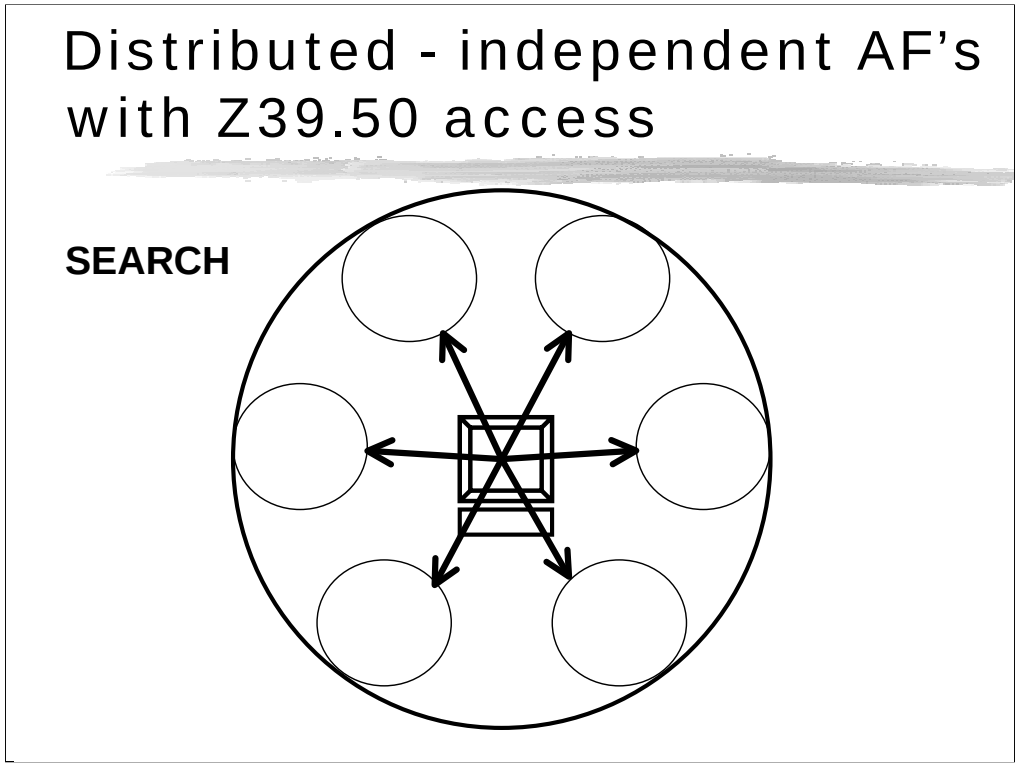
Tag	I1	I2	Subfield Data
010			ǂa n 80050515
035			ǂa (DLC)n 80050515
040			ǂa DLC ǂc DLC ǂd DLC ǂd NIC
100	0		ǂa Confucius
400	0		ǂa Konfuzius
400	0		ǂa K‘ung Fu-tzu
400	0		ǂa Kongzi
400	1		ǂa Kong, Qiu
400	0		ǂa K‘ung-tzu
400	1		ǂa K‘ung, Ch‘iu
400	0		ǂa K¯oshi
400	0		ǂa Konfu^t`si˘i
400	0		ǂa Kongja
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 HKCAN

{The Cyrillic text box is in True Type WP Cyrillic A.} This is not the 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. Actually with Unicode the roman script diacritics would appear with the letter rather than before the letter shown here, but this just gives you an idea of what it would be like.

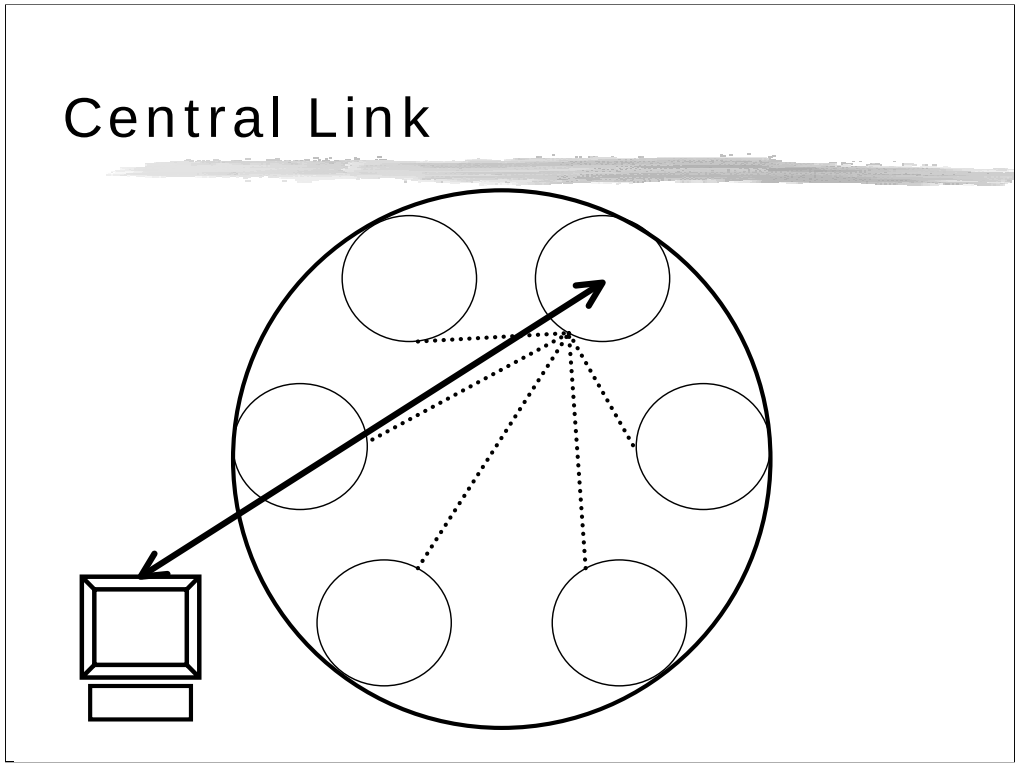
There is no particular order to the arrangement of the references, except to place the non-roman scripts following the roman scripts. This model shows English, German, Italian, Chinese, Japanese, Korean, Russian, and transliterations (including Wade-Giles and pinyin for the Chinese, since the Library of Congress just switched to use pinyin).

Notice also the new MARC 21 capability to include the URL for a Web page in the last 670 note field.

This also shows the use of a linking 700 field to show that an authority record was located at HKCAN and the form of authorized heading according to your rules. I know in HKCAN you use the 7XX fields for the authorized form in the alternate language – English or Chinese - and are able to use this information in OPAC displays to direct users to additional material cataloged under that alternate form. That enables bibliographic control for collocation under the name of the person or corporate body. Your authority records could potentially be a wonderful resource for a future virtual international authority file to assist with end-user displays in the language and script they prefer.

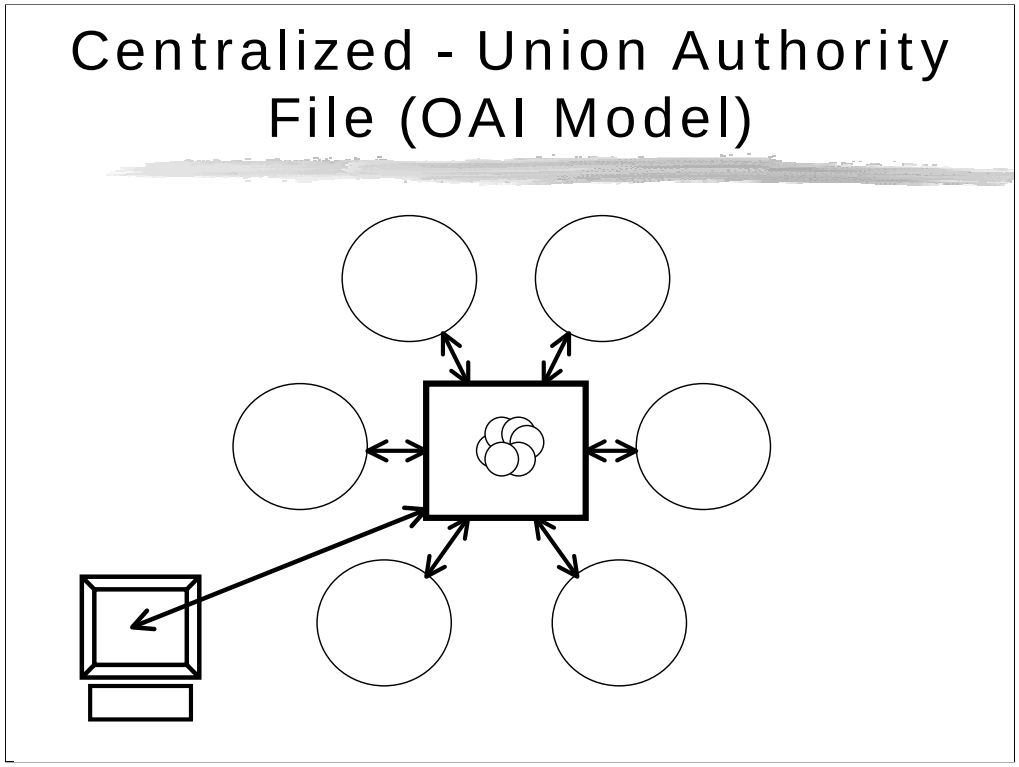


There are several models for a virtual international authority file. For a distributed model, a searcher would use a standard protocol like the next generation of Z39.50, to search the independent authority files of participating National Bibliographic Agencies or regional authorities.

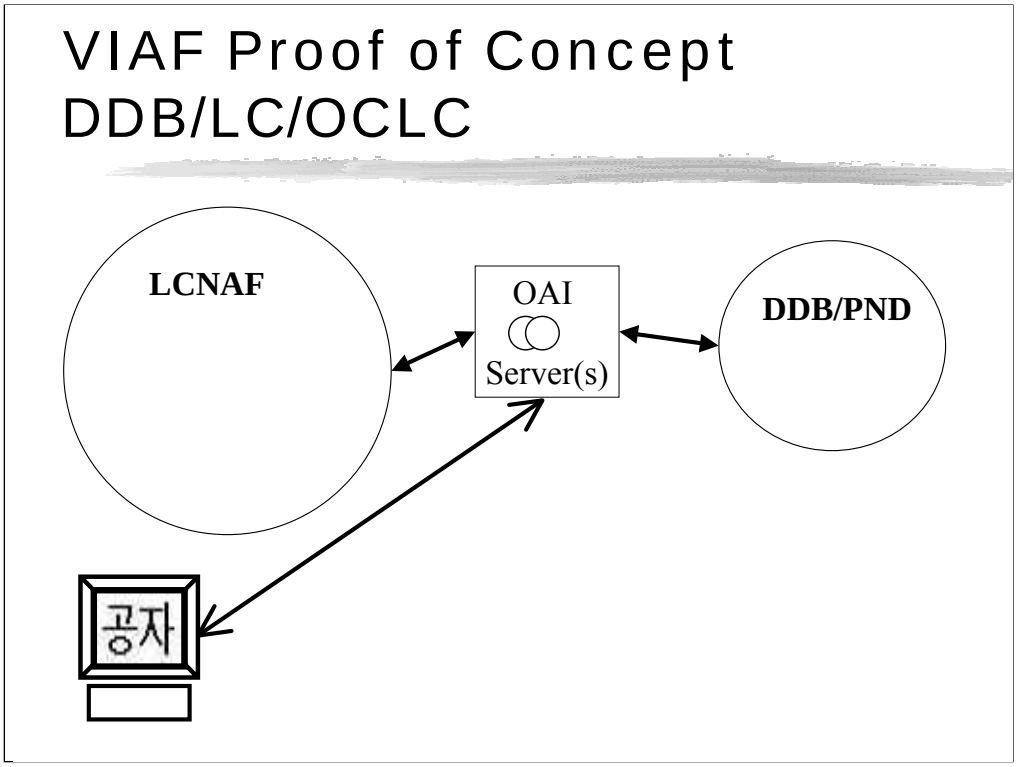


Another model is to have one central authority file and link all others to that, so that work would not need to be done by each national bibliographic agency with all other participants in this international universe. A cataloger would then get access to all the authority records for that entity worldwide 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.

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.



This shows a centralized model:
We may find that this model is the best approach in terms of record maintenance - The Open Archives Initiative (OAI) protocol model uses a server with harvested metadata from the national authority files. That information is refreshed in the 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). Unless we also build in the linking, we possibly will lose a level of precision in the searching in this model; but there are ways to include the links for entities in this model, too. There are many variations of models we could imagine.



The German National Library (Die Deutsche Bibliothek) and the Library of Congress together with OCLC have started a proof of concept project to test the centralized union authority file model using OAI protocols. The first stage of this project begins this year to link our existing authority records for personal names, testing the matching algorithms to see how much a computer can do for us and how much will require human intervention for matching and checking. It is hoped if this proves successful it can be the basis for a true Virtual International Authority File.

Matching retrospective files

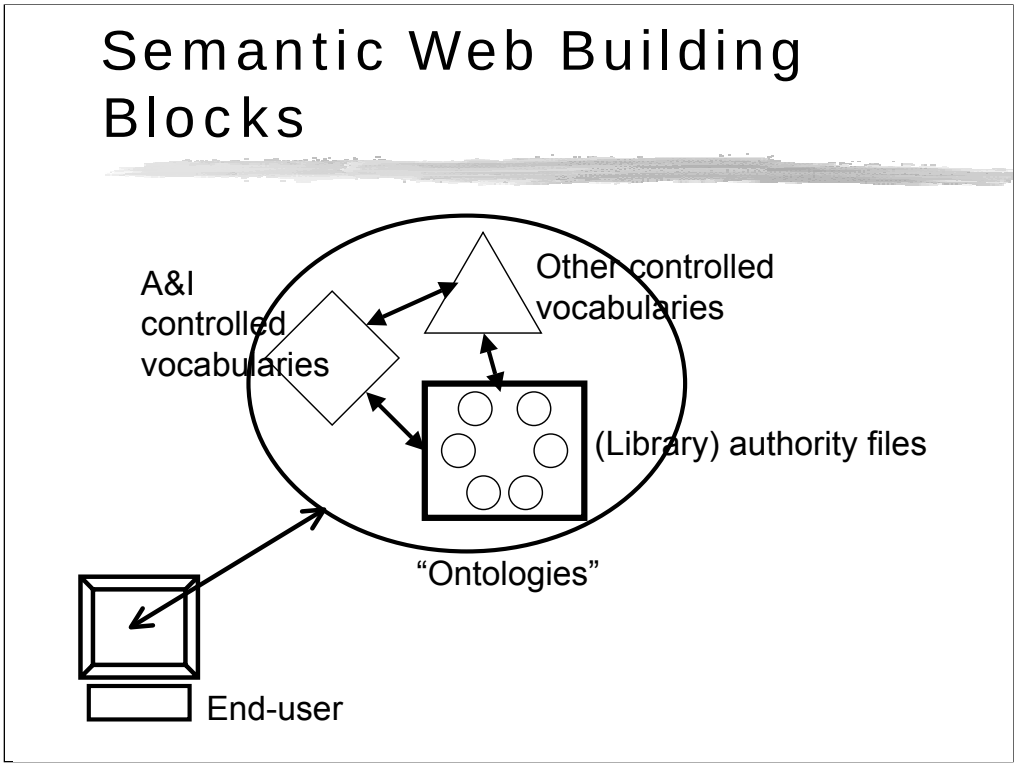
- One time project
- Matching algorithms, such as those developed by Ed O'Neill and others at OCLC
- Links
  - Text strings
  - Control numbers

If we agree that sharing authority information on a global scale is worthwhile, how do we get there?

Several major authority files exist, built according to their own cataloguing rules and rule interpretations. We need a one-time project to link the existing records for the same entity - a retrospective matching project. One suggestion has been to use matching algorithms, such as those developed by Ed O'Neill and others at OCLC, building on bibliographic clues for machine matching at a fairly high level of accuracy. A "proof of concept" project to test this approach is underway between OCLC, the Library of Congress, and the Deutsche Bibliothek (German national Library) in Frankfurt, Germany.

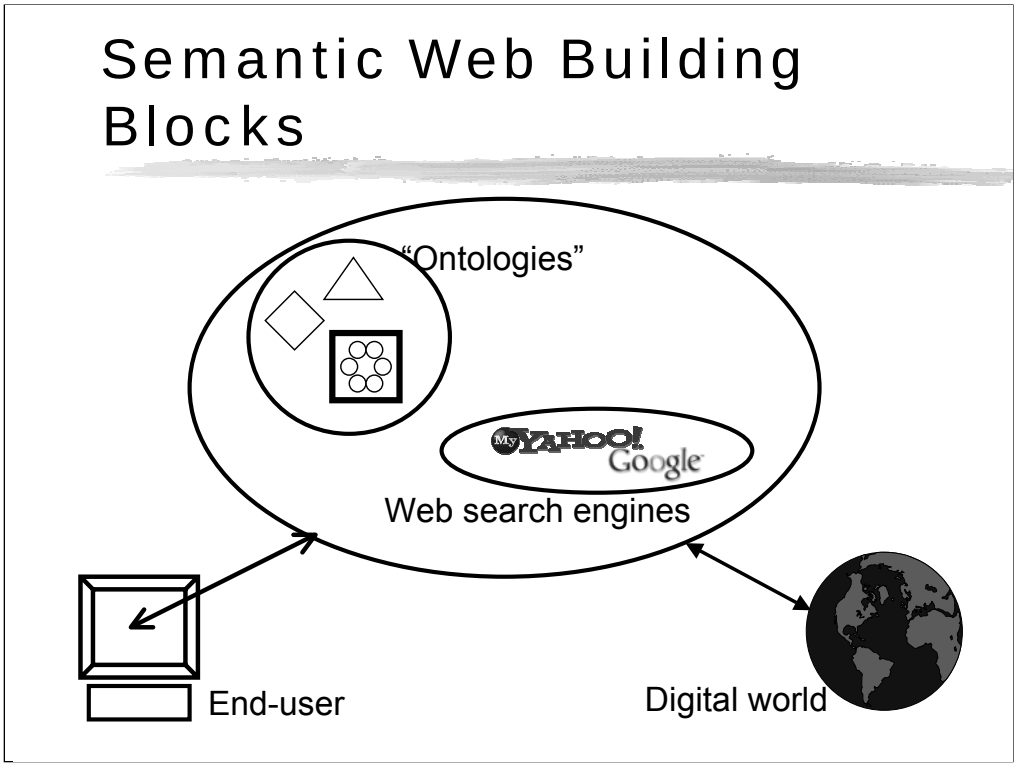
We would still have manual matching and checking to do, but expect machine matching will be a great help.

We could also have the computer add linking text strings and record control numbers or an entity identification number to facilitate later links and pathways to preferred forms for displays.



We can also 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.

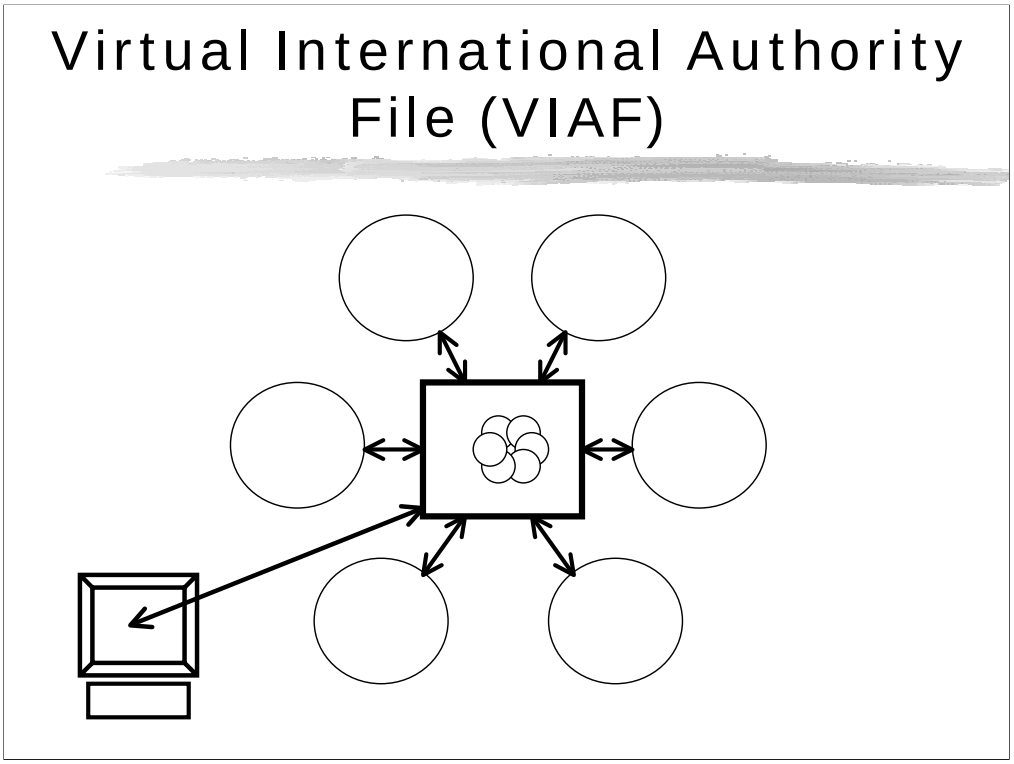


You can see that we would also build in the search engines and future tools that as a collective resource would connect us to the entire digital world.

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 - we are definitely talking 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.



We are preferring this model for now, as it seems to hold the most promise for scalability – to include the connections to all the major authority files worldwide.

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

LC and DDB have made the start in linking our authority files for personal names. We signed a memorandum of understanding in August 2003 in Berlin. Once we prove this model, we would very much like to test adding other authority files, but remember our goal is to make this information freely available to users worldwide. Great challenges are ahead but the technology now makes this test possible.

Comments or questions?