VIAF Guidelines

Adopted 14 August 2015 in Cape Town, South Africa
Revised 16 July 2019

Aim and scope of this document

This document gives a short description of what is needed to successfully participate in the Virtual International Authority File (VIAF®), including: what records should be submitted, basic requirements for records, where VIAF looks for information, how you can influence the clustering, and common problems.

Outline:

- Key principles
- Types of authorities currently supported
- Record formats
- Minimum information required
- What types of records to include in VIAF
- Information by types of entities
- Typical problems
- Forced links
- Reassigning authority control identifiers
- Where VIAF looks

The intended audience is VIAF participants and applicants.

Key principles: The more complete and accurate the information, the better the matching

- VIAF reflects the data submitted to it by the participants. VIAF does not create data but only processes data submitted by VIAF participants.
  - The participants are encouraged to submit all their authority data corresponding to the VIAF scope (see below).
  - VIAF respects the editing rules set up by each participant in building authority records.
- VIAF is periodically updated and reprocessed to be compliant with updates its participants submit to it. While the VIAF team at OCLC does its best to properly match and represent an institution’s records in VIAF and regularly keeps the participants informed on the processing of their data, the participants are encouraged to pay attention periodically to how their data appears in VIAF and let OCLC know about problems.
- VIAF ingests data from both authority and bibliographic records. VIAF uses information from associated bibliographic records to enable the matching process and enhance its quality.

- The authority data is public, along with information extracted from the bibliographic records and stored in the VIAF “processed” versions of the source records. The VIAF clusters, and consequently data within them, are under an open-data license (ODC-by).
- VIAF can match bibliographic records to authorities either by explicit identifiers, string matching, or both.
- VIAF’s intended scope includes all types of library authorities except for topical subjects.
- Updates to authority and bibliographic data should be sent to VIAF regularly so that the VIAF database accurately reflects a participant’s data. VIAF participants can decide how frequently they wish to send updates (e.g., weekly, monthly, yearly).
- Contact information for the person responsible for data quality issues should be accurate, as we may reach out to this person for problems reported to us by VIAF users or issues we encounter in the participant’s data.

Types of authorities currently supported

- Personal
    - Includes families
- Corporate
    - Includes conferences
- Works
- Expressions
- Geographic
    - Includes jurisdictionals

Record formats

Data must be expressed in OCLC-parsable syntax. OCLC has routines in place to handle various flavors of MARC 21 and UNIMARC in ISO 2709 or XML.

Character encoding should be in Unicode (preferred) or MARC-8.

Data delivery

GZIP is the preferred method if the data is compressed.

The data should arrive in two files: one for authorities and one for bibliographic data. If either file exceeds 2G, then it is acceptable to send records in multiple files, but never one record per file.

OCLC has an ftp site where data can be deposited, or we can do OAI harvesting. If needed, harvests can be pulled from the source site via ftp or http.

Updates should include adds, changes, and deletes. Please inform OCLC if your update file is a replacement of the full file.

Please notify OCLC when updates are available. You can send an email to OCLCVIAF@oclc.org, or we can schedule a data pickup on the same day of every month.

Minimum information required

- A stable local identifier associated with all authority and bibliographic records, typically found in the 001 fields of MARC records
- One or more “preferred” forms of a name associated with the entity
- Enough well identifiable information to differentiate from or match to records from other sources
- One record/entity, except in cases in which the source has identifiable parallel records (e.g., the French and English records in the Canadian file)

What types of records to include in VIAF

Typically, VIAF expects participants to send relatively complete records, but VIAF aims for completeness as well. In general, if an institution makes a record publicly available, OCLC would like to have it in VIAF. Currently by default, VIAF eliminates provisional records from UNIMARC and MARC 21. If an institution contributes bibliographic information to OCLC’s WorldCat®, OCLC may be able to use that for the bibliographic records that VIAF needs.

Information by type of entities

Generally useful information for all types of entities

- Identifiers
  - VIAF, ISNI IDs
  - ISBNs, ISSNs
  - Other authority record IDs, e.g., GND or LC
- Associated titles of works
  - Make sure that VIAF is aware of how to link your bibliographic records to authorities.
  - Tell us how to recognize titles embedded in authority notes (e.g., 670s).
- Alternative forms of names
- Associated names
- Non-Latin versions of names and titles

Persons

- Birth, death, flourished dates when available and when it is possible to make them public
- Titles are especially important to embed in authority records if no bibliographic record is associated with the entity.

Works and expressions

- When VIAF finds an expression record (e.g., a record describing a translation of a work), but no equivalent work record, it will generate the work record and link it to the expression.
- It is very helpful to code the roles of various entities (persons and corporate bodies), e.g., translator, illustrator, performer… .

- Original script is preferable when possible, e.g., Chinese characters for Chinese titles. Transliterations are also useful in matching.

Geographics

- VIAF treats geographics as a separate type; however many, if not most, geographics in VIAF fall more into jurisdictional names, such as names of cities. These sometimes overlap with corporate names, so it may occur that both geographic and corporate names coexist in a VIAF cluster.

Typical problems

Here are some of the more common problems VIAF encounters clustering names. See the Forced Links section below for ways to cope with recalcitrant clustering problems. VIAF contributors are encouraged to have a particular look in the following cases.

- Mixed names (mixed homonyms)
    - Mixed names often happen when titles that should be separately associated with different entities get all linked to one of these entities. This leads to merging unrelated entities, causing problems for everyone.
    - If the record is known or suspected to be undifferentiated (mixed names), this needs to be indicated.
- Missing titles, dates
    - VIAF needs enough information to differentiate the entities you are describing. Personal names with no titles or dates associated with them (either from bibliographic records or from within the authority records) are difficult to match correctly; however, if the name appears unambiguous, matching is possible.
- Unusually encoded information
    - If an institution follows some local convention that is not typical (for example, how it differentiates between cities of the same name), please let OCLC know your conventions so it can try to utilize the information with profit.
- Differentiation using language-specific information
    - In personal names, this typically is in a MARC 21 $c subfield as qualifier in the cataloging language (e.g., “Dramatist” in English vs. “Dramaturge” in French). For most matching, such information will be ignored because so much variation is seen.
- Duplicates
    - Duplicate records within a file cause problems because, in general, VIAF tries not to include more than one record from each source in a cluster. Duplicates often result in clusters being split.
    - VIAF periodically reports records that look like duplicates. This can be useful, both in helping an institution to eliminate duplicates in its file and, at least as important, to identify problems in VIAF pulling together records that it should not, e.g., by ignoring some carefully coded information that an institution supplies but that OCLC is not aware of.
- Classical names, kings, queens, popes

    - These can be very difficult because of the wide variety of forms of names in various languages. Report problems, and OCLC will manually correct these.

Forced links

VIAF has several ways to either force or at least strongly recommend links between source records, and therefore affect the clustering.

Before doing explicit forcing, it is often enough just to ensure that an additional title is added or an alternative form of the name. This is especially helpful if it looks as though VIAF completely missed considering the match at all (996s in the processed record are possible matches that VIAF rejected; 998s are actual links made between source records).

It is important to understand how VIAF views the clustering process. VIAF first matches each record in each source to all the records in all the other sources. Once those source-record-to-source-record links have been made, VIAF divides the records into clusters based on the links.

So, one way an institution can affect the clustering of its records is to include the ID of a source record from someone else that is thought to be a match, such as an LCCN or BNF ID. This linkage is considered a very strong indication that the two records should reside in the same cluster.

Including VIAF IDs in records: When VIAF sees that a VIAF contributor includes a VIAF ID in its records, VIAF looks at the current production database and changes that VIAF ID to a series of links to all the source IDs in that cluster. The clustering then proceeds as normal with the addition of those strong links. Including VIAF IDs in records is the equivalent of putting strong links to each record in the VIAF cluster. While useful, putting VIAF IDs in records effectively points to all the other records in the cluster, which can lead to problems if one of them is there in error.

The best practices are:

1) When possible, point at only one record from one VIAF contributor (for instance, when data coming from only one VIAF contributor is used as a source for establishing the authority record), to point at this particular source ID.
2) In other cases, to pay a strong, particular attention to the validity of clustering before citing any VIAF ID. It leads to problems with questionable clusters.

It is possible for VIAF to ignore explicit links when there is a conflict. Links from xA and ISNI have the highest precedence.

- **xA record (Extended VIAF Authority):** When VIAF is unable to algorithmically match some of the source authority records with each other, they can be manually pulled together into a single cluster using an internal table.

Finally, an institution can contact OCLC about a bad cluster. This is especially important if it looks as though there is a systematic error you think VIAF is doing that may affect many clusters. OCLC can create an xA record to pull otherwise unrelated records together as well as two xA records that will pull a cluster apart. The ISNI Quality Team is able to indicate the same sort of information when correcting ISNI records, so that information should flow back to VIAF to help it as well.

Reassigning authority control identifiers

If your authority identifiers will change, please contact us beforehand to discuss the logistics of reloading your entire authority and bibliographic file to VIAF. The old authority identifier needs to be included in the new authorities records for us to match the identifiers. We recommend that the old identifier be placed in a MARC 21 or UNIMARC 035 authority field.

Where VIAF looks

- Alternate forms of names
  - MARC 21 4XX
  - UNIMARC 4XX and 7XX
- Associated names (used for cross-references)
  - MARC 21 authorities 5XX, 7XX
  - UNIMARC authorities 5XX, 7XX
- Information about the headings:
  - UNIMARC coded subfields in 2XX, 4XX, 5XX, 7XX especially:
    - $5 relationship code (with special values for pseudonym, married/maiden name, name in religion, acronym, etc.)
    - $7 script of cataloging and script of the base access point
  - MARC 21: $4 in 4XX and 5XX
- Country of publication
- Biographical dates
  - MARC 21 authorities 046
  - UNIMARC authorities 103
- Differentiated/Undifferentiated record
  - MARC 21 personal authorities 008 byte 32
- Forced, suggested links
  - VIAF IDs
    - MARC 21 authorities 024
    - UNIMARC authorities 810 $a
  - Source IDs (the VIAF contributors are encouraged to check whether the metadata associated to the source IDs they provide in their own record is curated)
    - ISNI
      - MARC 21 authorities 024
      - UNIMARC authorities 010 ($a valid ISNI, $y cancelled ISNI, $z erroneous ISNI)
    - LCCN
      - MARC 21 authorities 010
      - MARC 21 700 $0
    - Source IDs as VIAF processed records, ORCIDs, ISNIs
      - 670 $u e.g., http://viaf.org/processed/DNB%7C118505173
      - MARC 21 authorities 024
  - Miscellaneous

- VIAF looks for parallel records in LAC and NLB.
- Links from xA are treated as forced.
- Links from ISNI reviewed records are treated as forced.

- Gender/Sex
    - MARC 21 375
    - UNIMARC 120
- GeoNames
    - MARC 21 034 $a
    - MARC 21 751 $0
- ISBN
    - MARC 21 bibliographic 020
    - UNIMARC bibliographic 010
- ISSN
    - MARC 21 bibliographic 022
    - UNIMARC bibliographic 011
- Language
    - MARC 21 authorities 377
    - MARC 21 bibliographic 008 bytes 35 – 37, 041
    - UNIMARC authorities 101
    - UNIMARC bibliographic 101
- LCCNs (see also Forced Links)
    - MARC 21 bibliographic 010, 050
    - DNB bibliographic 036
    - UNIMARC bibliographic 680
- Nationality/country associated with a person
    - MARC 21 authorities 370
    - UNIMARC authorities 102
- Name as subject
    - MARC 21 authorities 008 byte 14=b
    - UNIMARC authorities 106
- Names from statement of responsibility
    - DNB Authorities 692
    - MARC 21 bibliographic 245
    - UNIMARC bibliographic 200
- Occupation
    - MARC 21 authorities 372, 374
- Provisional records (VIAF ignores them)
    - MARC 21 authority leader byte 17 (encoding level) value o (incomplete) or 008 byte 33 (level of establishment) value c (provisional) or d (preliminary)
    - UNIMARC authorities 200 $a byte 8
- Publisher
    - MARC 21 bibliographic 264 1 $b (RDA records) or 260 $b (pre-RDA)
    - UNIMARC bibliographic 010, 210
- Publication dates
    - DNB authorities 692

- Titles
  - DNB authorities 692
  - MARC 21 authorities 670
  - MARC 21 bibliographic 245, 210, 130, 240, 243, 242, 246
  - UNIMARC authorities 810
  - UNIMARC bibliographic 200, 4XX, 500, 501, 510, 512, 513, 514, 515, 516, 517, 518, 532, 540, 541
- Non-Latin character forms are encouraged.
- VIAF IDs (see Forced Links)