Description of the VIAF (Virtual International Authority File) Dataset

Thomas B. Hickey, Chief Scientist, OCLC Research

Jeffrey A. Young,, Software Architect, OCLC Research

Name of dataset

VIAF (Virtual International Authority File)

Brief description of dataset

VIAF virtually combines multiple library authority files into a single name authority service. The system mines and clusters variant names for a given entity (chiefly, persons and organizations), links the corresponding source records, and assigns a URI to each cluster. The dataset is built using advanced algorithms developed by OCLC Research, a global leader in applied research related to library information, and is the product of an ongoing collaboration of OCLC and a group of national libraries, other leading libraries and other cultural heritage organizations. Available as Linked Open Data (LOD), VIAF is leveraged by freebase.com and an expanding array of other agencies and services.

URL

http://viaf.org

VoID (Vocabulary of Interlinked Datasets) description

http://viaf.org/viaf/data

Version date and number

Updated monthly

Licensing

The VIAF dataset is made available under an Open Data Commons Attribution License (ODC-By)

To facilitate interoperability, OCLC strongly encourages the use of VIAF URIs in all appropriate circumstances. The canonical structure of a VIAF URI is http://viaf.org/viaf/[numerical value] (Example: http://viaf.org/viaf/49224511)

Adherence to ODC Attribution instructions for the correct assertion of attribution is encouraged. The preferred form of attribution for VIAF is:

> "This [title of report or article or dataset] contains information from VIAF (Virtual International Authority File) which is made available under the ODC Attribution License."

Special cases: In circumstances where providing the full attribution statement above is not technically feasible, the use of canonical VIAF URIs is adequate to satisfy Section 4.3 of the ODC Attribution License.

Datasets that integrate the VIAF dataset directly should still make an effort to incorporate the standard attribution text somewhere in relation to their dataset description when appropriate.

Availability

VIAF is available from a public web site (http://viaf.org) written using a public API, plus bulk downloads. The API is documented at http://oclc.org/developer/services/viaf.

Topic coverage

Entities whose names are controlled by libraries, including people, corporations, and places, as well as FRBR (Functional Requirements for Bibliographic Records) works and expressions. (FRBR is a model of bibliographic relationships developed by the global library community).

Sources of the data

Library authority files from a number of national libraries, plus selected regional and trans-national library agencies.

Purpose

To lower the cost and increase the utility of library authority files.

Method of creation

The source authority files and associated bibliographic metadata are either sent to or harvested by OCLC, usually on a monthly basis. Metadata about entities is extracted from bibliographic records supplied by contributors directly or from the WorldCat® database (the world's largest aggregation of library catalog data, hosted by OCLC) and merged into the authority records. These enhanced authority records go through a matching process, ambiguities are resolved and clusters built from matching records. Record IDs are maintained across updates with redirects in the Web service API when clusters get merged.

Maintenance

The file is rebuilt each month using the latest source data.

Web browser usage

The HTML pages use Google Analytics to collect statistics on use. Each month VIAF receives about 55,000 visits and 260,000 page views from 30,000 visitors.

Web API usage

The web service currently receives about 10 million searches/page accesses each month, many of which appear to be from commercial web harvesters.

RDF usage

VIAF receives about 40,000 accesses to the RDF view of clusters each month. Again, many of these appear to come from commercial harvesters.

Bulk Distribution

OCLC makes the VIAF dataset publicly available under the ODC-By license. Details of the dataset and licensing are available in the VoID description referenced above.

Size of the Dataset

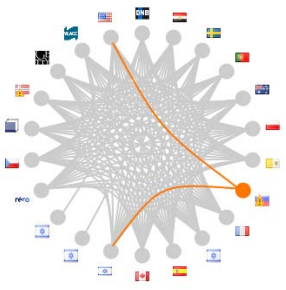
The dataset is built from information supplied by nearly two dozen files consisting of 25 million authority records and 110 million bibliographic records. VIAF builds some 20 million clusters from those 25 million source records, finding about 24 million relationships between source records. The 20 million clusters and the enhanced authority records are visible through the HTML interface, but only the clusters are harvestable and made available through bulk download.

External Connectivity

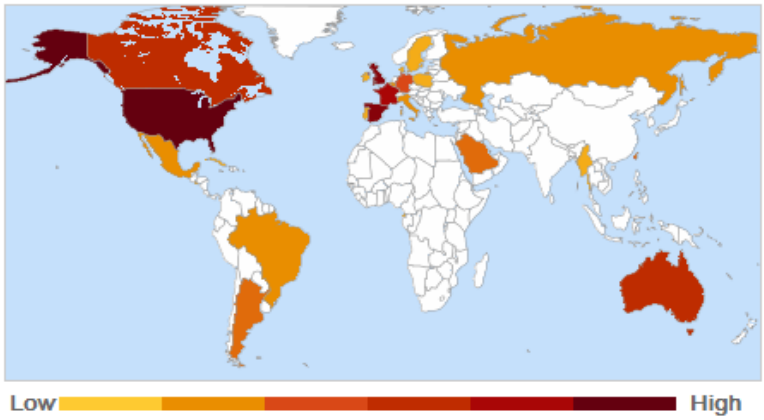
The RDF and HTML views of the data connect to the contributing source institution's web service whenever possible. So, in addition to the 24 million equivalences created by the clustering, VIAF makes 32 million *owl:sameAs* and 8.5 million *skos:exactMatch* links, including 300,000 links to DBPedia.

Examples

The HTML display in the public interface at http://viaf.org displays a summary of the information contained in each cluster. An interactive diagram showing how matches between records used to establish the cluster is shown, along with the various forms of the name:



Other visual displays of information, such as a publication timeline and map, are also available, e.g.:



The evolution of data modeling and linked data in VIAF

The VIAF RDF model has evolved over time. The first iteration used SKOS almost exclusively. This seems reasonable in terms of incoming authority files which are primarily concerned with controlling the name of things (via *skos:prefLabel* and *skos:altLabel*) within a limited domain (e.g. national cataloging practices). From a global clustering viewpoint though, the emphasis isn't so much on imposing labels across all domains as it is on recognizing, identifying, and relating "the thing" that all these contributors are trying to name. Nevertheless, the first iteration of VIAF modeled "the thing" as a *skos:Concept* despite the fact that there was no *skos:prefLabel* selected from the list of competing possibilities.

Sometime later, Friend of a Friend (FOAF) was gaining popularity and there was a growing sense that "the thing" driving the clustering wasn't merely a *skos:Concept* (e.g. an abstraction like "Hunger") but was more like a *foaf:Person* or *foaf:Organization* instead. Thus in the second model iteration "the thing" was described using FOAF terms instead of SKOS.

Although this approach seemed more realistic in terms of describing "the thing", FOAF didn't seem well suited for establishing links back to the controlled vocabulary aspects of contributed records. There was a sense we should switch back to SKOS, but then the idea of doing both was considered. Coincidently, the Resource Description and Access (RDA) community had developed a library-oriented RDF vocabulary that had terms analogous to FOAF like *rdaEnt:Person*, *rdrEnt:CorporateBody*, so those terms got thrown into the model as well. It wasn't completely clear, though, whether "the thing" could be a *skos:Concept* and a *foaf:Organization* and a *rdaEnt:CorporateBody* all at the same time, so separate identifiers for each were created using *#skos:Concept*, *#foaf:Organization*, and *#rdaEnt:CorporateBody*. Although these hash URIs all piggy-backed on the same VIAF URI, there was no relationship asserted between them in the data.

The solution eventually revealed itself in the form of *foaf:focus*, which defines a meaningful relationship between one or more *skos:Concepts* (e.g. controlled vocabulary terms) and an *owl:Thing* (e.g. *foaf:Organization* or *rdaEnt:CorporateBody*). Using this insight, we were able to factor out the incoherent entities (e.g. the *#skos:Concept* aspect of "the thing"), merge the redundant *#foaf:Organization/#rdaEnt:CorporateBody* entities into the previously untyped "cluster" entity

identified by the canonical "VIAF URI". The result was a much more streamlined and intuitive hub-and spoke model illustrated and described in this blog post: http://outgoing.typepad.com/outgoing/2011/04/changes-to-viafs-rdf.html .

Known Shortcomings

- Matches between source entities is fairly conservative, so many possible matches are missed
- Limited ability for those outside the library community to correct errors
- Monthly updates mean changes made to the source files could take more than a month to show up in VIAF
- Geographic names that are not jurisdictional names are not yet included
- Differences in rules and conventions can cause difficulties. A common example is for one authority file to have entities for both *Samuel Clemens* and *Mark Twain*, while another only has *Mark Twain.*
- Several types of entities, such as works and expressions, have been added that haven't been accounted for in the RDF model yet.
- Schema.org has emerged as a popular alternative to FOAF and RDA, but hasn't been incorporated into the model yet.

URIs Referenced

VIAF Service: http://viaf.org

VoID (Vocabulary of Interlinked Datasets): http://www.w3.org/TR/void/

VIAF VoID Document: http://viaf.org/viaf/data

Open Data Commons Attribution License (ODC-By): http://opendatacommons.org/licenses/by/

VIAF API documentation: http://oclc.org/developer/services/viaf

FRBR (Functional Requirements for Bibliographic Records): http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

The Friend of a Friend (FOAF) Project: http://www.foaf-project.org/

foaf:focus: http://xmlns.com/foaf/spec/#term_focus

Simple Knowledge Organization System (SKOS): http://www.w3.org/2004/02/skos/

skos:Concepts: http://www.w3.org/TR/skos-reference/#concepts

OWL Web Ontology Language: http://www.w3.org/2007/OWL/wiki/OWL_Working_Group

owl:Thing: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/#owl_Thing

Resource Description and Access (RDA): http://www.rda-jsc.org/rda.html

Related Resources

IFLA Study Group on the Functional Requirements for Bibliographic Records. 1998. *Functional requirements for bibliographic records: final report* : K.G. Saur.

Young, Jeff. 2011. Changes to VIAF's RDF (Outgoing blog post, April 12, 2011) http://outgoing.typepad.com/outgoing/2011/04/changes-to-viafs-rdf.html.