Towards the creation of integrated authority files in the domain of science and technology: an Italian use case

Elena Cardillo, Iryna Solodovnivck, Maria Taverniti
Institute of Informatics and Telematics, National Research Council, Italy
{elena.cardillo, iryna.solodovnik, maria.raverniti}@iit.cnr.it

Introduction

Using shared application profiles such as controlled vocabularies and authorities is fundamental for the efficient knowledge organization and description of bibliographic resources. They provide trustworthy access points to bibliographic information resources driven by authority control which facilitates controlled access points for each entry (e.g., personal and corporate names, etc.). This enables semantic interoperability in bibliographic databases and digital repositories. In addition, different national and international organizations create, maintain and openly share their subject heading systems, classifications, thesauri, which if used will contribute to high accuracy, retrieval, interoperability and enhanced controlled access points to ‘common knowledge’. Nevertheless, sometimes there is the need to create local Authority Lists (ALs) and Authority Files (AFs) for niche knowledge areas not completely covered by existing KOSs. Following this perspective, this poster shows the creation of a local name AF that integrates ALs in the domain of Science and Technology for their application in the CNR S&TDL project\(^1\), where, together with other KOSs, they will be used as a classification schema organizing information resources at a general level.

CNUCE / CGI Corpus

A corpus of documents belonging to the University National Centre of Electronic Calculation (CNUCE), a CNR institute that in the 1970s and 1980s started the spread of Computer Science and the Internet in Italy\(^2\) and to the General Informatics Commission (CGI).

A corpus of 400 documents has been collected from CNR archive and personal archives of IIT-CNR researchers in Pisa and then digitized (resolution 600dpi) as pdf files:

- Digitization quality compromised by the poor quality of the original documents;
- Documents are mainly technical reports, constituent acts and memoranda of the two institutions.

Authority lists needed to provide a control for name variants and name disambiguation, as well as more detailed access to and better retrieval of CNUCE documents in the CNR S&TDL Digital Library.

Authority list Construction and Integration

Term Extraction and NER

Person/Organization Authority lists

VIAF\(^5\) and Online Resources

Results

Significant effort was needed in order to review and match different variants of the same Name: e.g., Biorci prof. Giuseppe, Giuseppe Biorci, Prof. Biorci, G. Biorci, etc... CNUCE documents became publically available only a few years ago. This resulted in: a shortage of historical-literary studies, the absence of large reference works providing extended knowledge about the extracted personal names, and insufficient coverage of these names by external Authority files; many persons cited in the CNUCE documents are managers or administrative staff with no publications. This justifies the limited number of mappings to VIAF, as shown in the Table (about 60% for Persons’ names, and just 27.5% for Organizations’ names.

<table>
<thead>
<tr>
<th>Type</th>
<th>Total</th>
<th>Municipal</th>
<th>Matching to VIAF</th>
<th>Data from Bibliography</th>
<th>Data from other sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authority List Person</td>
<td>257</td>
<td>700</td>
<td>620</td>
<td>280</td>
<td>418</td>
</tr>
<tr>
<td>Authority List Organization</td>
<td>295</td>
<td>1000</td>
<td>110</td>
<td>205</td>
<td>395</td>
</tr>
<tr>
<td>Total</td>
<td>552</td>
<td>1,100</td>
<td>930</td>
<td>485</td>
<td>813</td>
</tr>
</tbody>
</table>

Conclusions

Authority lists have been used for the indexing of CNUCE documents then ingested within the CNR S&TDL Digital Library, easing the access to data related to CNUCE within the platform. Benefits derived from the creation of integrated Authority lists for CNUCE are: 1) the possibility to reuse a consolidated classification, identification and discoverability of people and connected topics; 2) the possibility to view name records as established by other nations thanks to the mapping to VIAF, thus making the authorities truly international and facilitating research across languages through interrelated identifiers anywhere in the world. CNUCE subject headings list are in the process of being created on the basis of the list of keywords extracted by T2K with mapping to the Nuovo Soggettario\(^7\) and to and Library of Congress Subject Headings (LCSH):\(^7\) 1,250 relevant terms extracted from the CNUCE Corpus, resulted in 850 terms after validation. The formalization of CNUCE Authority lists in RDF is a work in progress to allow their publication as Linked Open Data in a data hub (e.g. CXN DataHub).

References

1. Science & Technology Digital Library. Available at: http://stdl.cnr.it/
3. VIAF. The Virtual International Authority File. Available at: http://viaf.org
5. Solodovnivck, Iryna. Authority files and VIAF. Available at: http://csd.unige.it/dai/5th/Solodovnivck