

## **LEAF**

**Pierre Clavel**  
Swiss National Library  
Bern, Switzerland

### **Introduction**

Name authority files today can be divided into two kinds:

- An independent database, not linked to a catalogue, whose purpose is exclusively to record the different forms of the names of persons and/or corporate bodies. The *Personnamendatei* maintained by Die Deutsche Bibliothek is an example.
- A table (or group of tables) in the database of a library system, whose added purpose is to bring heading consistency to the bibliographic records with links from the bibliographic record to the authority. Its authority records may have been derived from an authority file of the first kind.

This second kind has been commonplace in library systems since their second or third generation. Despite the development of library networks and virtual union catalogues, these authority files have remained tied to a single catalogue. The first reason for not integrating a single authority file to several catalogues is a question of performance: gathering full records from different databases in real time is already demanding in terms of network resources and processing power; to add the requirement of building up records for display from links to their distant headings would be even harder.

Moreover, it is not certain that the market for a central authority file would be large enough for vendors to recoup their development cost. Libraries often prefer keeping a local database in addition to joining a large central union catalogue. They like to stay able to add information of their own, such as subject headings or notes, to a downloaded bibliographic record, in order to tailor their catalogue to their local users' (perceived) needs. The language of the downloaded information is one of the justifications for doing so, when it is different from the targeted users' one.

Authority records are not different in this respect (despite the concept!), since choosing a form over the others can be directed by language-related rules or, in some cases, be as subjective as adding subject headings and notes.

This freedom for individual catalogues in choosing the adopted form of a name has a drawback, now that distributed searches via Z39.50 expand. If a name form is adopted in a first catalogue but not in a second one, a search with this name will accurately yield records only if it is present as a rejected form in the second authority file and if the second target system is able to silently do a double-step linking. These conditions are seldom met.

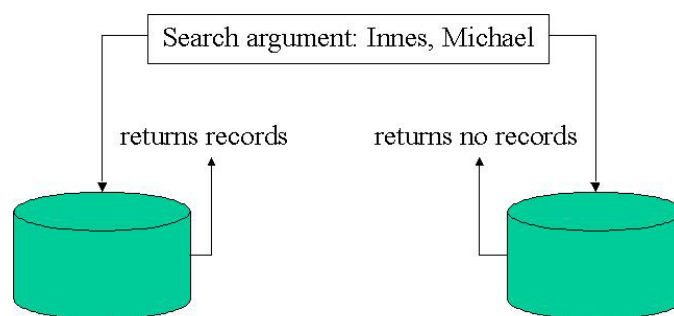


Figure 1: Possible problem with a distributed search

Headings given to records describing manuscripts face a further challenge. The identity of a person related to a manuscript is much more difficult to establish and the information about names is often sketchy. In addition, more often than one would think, the manuscripts relevant to a person (who may be its author, the topic, or the addressee of a letter) are widely scattered in various institutions, each having access to a different piece of information. By sharing as much information as available, archives can collectively identify people as if putting together the pieces of a puzzle. The need for, and benefits from, a shared name authority file are clear here, and this is precisely what the LEAF project is about.

## The LEAF project

The project LEAF (Linking and Exploring Authority Files, <http://www.leaf-eu.org>) started in March 2001, and has brought together fifteen organisations<sup>i</sup>. It is co-funded within the Information Society Technologies Programme of the Fifth Framework of the European Commission. It has obtained excellent reviews and aroused wide interest in the archival and library worlds. About 30 observing partners follow its progress and provide additional advice.

In an early stage of the project, it was hoped to make a virtual central authority file, the system searching via Z39.50 the authority files of the data providers and establishing links in real time. First simulation tests have shown that this was unrealistic from a performance point of view. LEAF is developing instead a model architecture for establishing links between uploaded authority records, without challenging the existence of local authority files, nor their differences of languages, formats, and names. This causes data redundancy and therefore updating issues, but was deemed acceptable with regard to performance. Eight different languages are spoken among the LEAF partners and are, above all, present in their data. Beside the availability of user interfaces in several languages, the project faces two multilingual challenges. The first one is in language-related differences of handling names. The order of components in compound names, the placement of particles and the way of numbering kings, popes and other homonyms are just a few classical examples. The second challenge is in exploiting language-dependent information. The addition of academic and nobiliary titles is generally standardized enough to be properly dealt with. The specification of activity or profession, be it to differentiate homonyms or provide more information, is much less standardized and using it in a cross-language way can be very difficult.

LEAF innovates in five aspects of the exchange and use of authority records:

- the exchange format, EAC (Encoded Archival Context)
- the linking process
- the integration into a distributed search service
- the usage-driven improvement process
- the addition of annotations

## Format

Computerized archival descriptions are globally less standardized than bibliographic records. Archival standards traditionally refer more to content and content organisation than format. There has been however a big step forward in 1998, with the introduction of EAD (Encoded Archival Description, <http://www.loc.gov/ead/ead.html>), an XML DTD that has been widely and quickly adopted throughout the world for archival description.

EAC is another XML DTD (<http://www.library.yale.edu/eac/>), parallel to EAD with which it has strong structural links. It allows the provision of authority information about persons, families and corporate bodies. Whereas EAD is clearly designed for archival description, EAC's objective and features make it a promising way for archives, libraries and museums to share data.

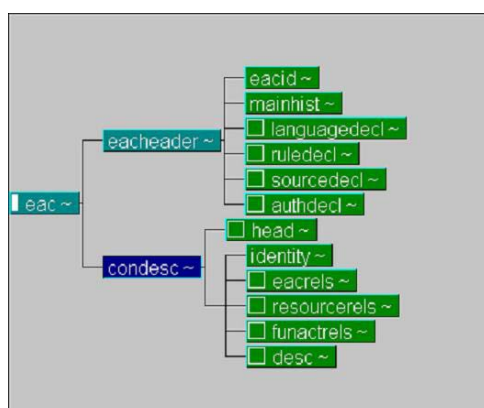


Figure 2 : structure of EAC, 1<sup>st</sup> and 2<sup>nd</sup> levels

LEAF is a co-promoter of EAC and contributes to its development. With its flexible granularity and nesting, XML can "absorb" a large range of different formats and function as a pivot between them. A set of tools converts authority records from the data providers' formats into EAC, either locally or on the central system. The LEAF system stores and processes records in this format and a set of style sheets allows their display and export in some of the formats of the data providers' systems, in addition to EAC itself.

## Linking

The LEAF database is first populated with the EAC representations of all authority records of the data providers, the *LEAF Authority Records*. This is done in several ways, depending on the technical environment of the data providers: the LEAF system can harvest an extraction of the local authority file via ftp or OAI, or

---

Z39.50 requests. The system then establishes links between them, when they are likely enough to describe the same person, following a set of *linking rules*. Linking rules use at least names from adopted and rejected forms and, when available, ID from a national or international authority file, as well as birth and death years. Further rules, such as exploiting activity information (e.g. profession), inter-script transliteration schemes, or usual variants in first names (e.g. Bob  $\approx$  Robert) are under consideration but might be left aside for the test phase.

When a link is established between two records, a copy of them is merged into a *Shared LEAF Authority Record*. A Shared LEAF Authority Record participates in the linking process of the records not yet examined. So, if a third LEAF Authority Record is found worth a link with a Shared LEAF Authority Record, a copy of the former will be added to the latter.

It is clear that the automatic linking process produces both noise and silence. Two records representing two different persons might nevertheless be automatically linked, because they do not contain enough discriminating information. On the other hand, two records representing the same person might not be automatically linked because they do not share an identical form.

In order to maintain consistence between local authority records and their image in the LEAF database, new and modified local records will be regularly uploaded in the same ways as the initial stock. During the automatic linking process that will follow, a modified LEAF Authority Record may have to be removed from a Shared LEAF Authority Record, or transferred into another.

## **Integration into a distributed search service**

MALVINE is a consortium running a distributed search service that accesses about a dozen of databases via Z39.50 (<http://www.malvine.org>). Its partners are also partners in LEAF. An interface between the LEAF and MALVINE systems allows expanding a search from one system to the other. For instance, an authority record found in LEAF reveals which institutions have that name in their authority files and, therefore, documents related to that person. A command can send out a request for document descriptions to the relevant databases, using the right adopted form for each of them. Conversely, from a document description found with MALVINE, one can switch to the LEAF Authority Record pertinent for a selected heading of the found record.

## **Usage-driven improvement process**

When a user searches in LEAF, retrieves a LEAF Authority Record or a Shared LEAF Authority Record, and requests all its content or uses it to search in MALVINE, the status of this record is changed to *Central Name Authority Record*. Despite a small risk of noise, this status indicates that a user was looking for this information and that the record has been used. [Z39.50 update?]. Data providers can see which of their records have been used and target accordingly the improvements they make to their data. Where appropriate, they could also derive hints for the development of their collections. On the other hand, data providers can assign to a LEAF Authority Record an *expiry date*, at which time it will be removed from the central database if it was never retrieved. The goal of this functionality, which will be used cautiously, is to avoid making a huge database full of names interesting no one.

## Annotations

The aim of annotation is to facilitate the exchange of information between users of all kinds.

Registered users as well as data and service providers can add annotations to the authority records of all types. Annotations may be temporary, persistent, or private.

*Temporary Annotations* allow users to provide an information about a LEAF Authority Record, be it independent or within a Shared LEAF Authority Record, subject to lead to a correction or an addition to that record. They trigger an automatic warning e-mail to the data provider(s) owning the original record(s).

*Persistent Annotations* can be attached to Central LEAF Authority Records, in order to provide a piece of information of general interest.

Registered users can save Central LEAF Authority Records in a private workspace provided on the server and add *Private Annotations* to them.

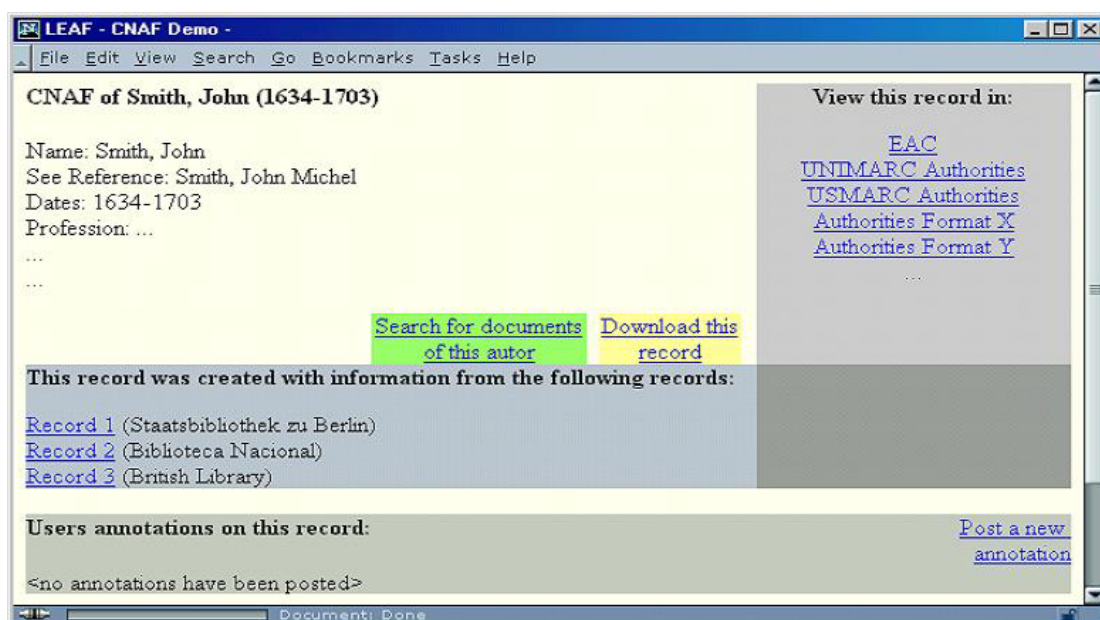


Figure 3 : LEAF record display interface prototype

## Next steps and conclusion

Internal tests have started in March 2003 with a few hundred carefully selected records from four partners. More than half-a-million records from further project partners and external organisations will be loaded in May to test the system in close-to-real conditions. Users outside the project will be invited to test the system from around mid-September, and the test phase will last until end November 2003. The project is due to finish at the end of February 2004, after which the project partners will set up a Consortium to scale up the system and run a full LEAF service.

LEAF offers a new perspective in a collaborative authority control, able to significantly increase the data quality of its members. Although it is primarily

designed to help end users and institutions concerned with manuscripts, it is also extremely valuable to libraries in general. Modifying imported bibliographic records can be reduced thanks to a consistency of headings at a higher level.

---

<sup>i</sup> COORDINATOR: Staatsbibliothek zu Berlin, Berlin, Germany

**PARTNERS:**

Biblioteca de Universidad Complutense, Madrid, Spain

Biblioteca Nacional, Lisbon, Portugal

British Library, London, United Kingdom

Crossnet Systems Limited, Newbury, United Kingdom

Deutsches Literaturarchiv, Marbach, Germany

Forschungsstelle und Dokumentationszentrum für Österreichische Philosophie, Graz, Austria

Goethe- und Schiller-Archiv, Weimar, Germany

Institut Mémoires de L'Édition Contemporaine, Paris, France

Joanneum Research - Institut für Informationssysteme und Informationsmanagement, Graz, Austria

Narodna in Univerzitetna Knjižnica, Ljubljana, Slovenia

Österreichische Nationalbibliothek, Vienna, Austria

Riksarkivet, Stockholm, Sweden

Schweizerische Landesbibliothek, Bern, Switzerland

Universitetet i Bergen - Forskningsprogram for humanistisk informasjonsteknologi, Bergen, Norway