Multilingual person names and Arabic script in authority files and research databases: the Mistara Project

Olga Andriyanova (BIMA), Asyeh Ghafourian (Bulac) & Benjamin Guichard (Bulac)

41st MELCom International Conference, Naples, 18-20 June 2019
Some recent history: Paris, 2002

- 2001: survey conducted by the Ministry of National Education on the French university library collections in non Latin characters
- Creation of the Languages and Civilisations Centre (Pôle des langues et civilisations) and the University Library for Languages and Civilisations Studies (BULAC): project in progress
- BnF and Sudoc catalogues preparing themselves for receiving bibliographic and authority records in original characters
- January 2002: report of the working group « Cataloguing of documents in non Latin scripts »
- May 2002: 24th MELCom International Conference at the National Library of France

National Library of France (BnF), François Mitterrand Site
Recommendations for Name Authority Data to be applied in the Sudoc catalogue
(on the basis of the Working group 2002 report)

Source: Cataloguing documents in non Roman scripts in SUDOC, the French university libraries network / Danièle DUCLOS-FAURE, MELCOM 2002)

- Data structure should include
  o at the record level: nationality, language(s) of the author (language of the country of origin; language of the title);
  o at the form level: script of the form; romanization standard;

- Form of authorized headings to be displayed in the bibliographic record: it is recommended to mention
  o form in original script;
  o romanized form, according to ISO 233-2 for Arabic;
  o French form commonly used;

- Variant forms: cannot be displayed in a bibliographic record, but have to be indexed so that they can be used for accessing the record
  o variants of the name;
  o names complying with American romanization standards (ALA-LC), if different;
  o names complying with older romanization systems, if different;
  o any other form of the name appearing on the document.
Vol. 9 No. 1 March 2003; Vol. 9 No. 2 June 2003 (Libraries and New Information Technologies)

- Round table « Le Fichier d’Autorité de la Bibliothèque de l’Institut du Monde Arabe (Base A’LÂM) » / Sarah Wassif
- Cataloguing documents in non Roman scripts in SUDOC, the French university libraries network / Danièle DUCLOS-FAURE
- Arabic Name Authority in the Online Environment: Options and Implications / Martha Speirs Plettner
- Towards an exchange network of Arabic authority records / round table report by Annick Bernard
- Bibliotheca Alexandrina and the Authority File / Manal Amin
- Working together: mailing list and discussion forum Arab authors (Auteurs arabes) launched (https://liste.cines.fr/info/auteursarabes)
The Mistara project

➢ Septembre 2018 : GIP BULAC (Public interest grouping BULAC) presents a research service project aimed at matching the existing onomastic databases for the Arab Islamic civilisation with the metadata produced by national catalogs, open and structured according to the web of data logic.

➢ Mistara (ruling board used by scribes) was chosen to name the project, in order to make an allusion to the process of the data alignment.
The Mistara project rationale: why is it relevant to revisit the question of authorities in Arabic script in 2019?

1. New bibliographic environment: ensure a better integration into the world of linked data and the semantic web
2. Assessment of existent authority records: poor record for certain types of authorities
3. Working together, with the right tools and with the right team, for a better future
4. Desire to ensure a quality service for the research community and for the library professionals
1. New bibliographic environment: ensure a better integration into the world of linked data and the semantic web

**Bibliographic Agency for Higher Education**

Sudoc catalogue **sudoc.abes.fr**

IdRef: authority database for Higher education and research **idref.fr** (authorities created in the frame of cataloguing of printed, electronic, manuscript resources and theses)

Triplestore **data.idref.fr** based on RDF data model

**National Library of France**

Catalogue **catalogue.bnf.fr**: allows search through bibliographic records, but also through name and subject authority data

Rameau **rameau.bnf.fr**: French national authority file (Répertoire d’autorité matière encyclopédique et alphabétique unifié)

Data **data.bnf.fr**: uses Semantic Web tools, ensuring a better visibility of the BnF data and linking them with external resources
In 2014, the two French bibliographic agencies (ABES & BnF) jointly expressed the French position on the RDA cataloguing code and launched the Bibliographic Transition Program in order to:
- FRBRise French catalogues;
- derive new cataloguing rules from RDA in RDA-FR: A French transposition of RDA;
- support all stakeholders in France through the changes induced by such a major normative and technological evolution.


French National Entities File (FNE):
- Synchronise the data managed by the two agencies in one database;
- Make the data relative to different entities (persons, agents, works, expressions, places, time spans, subjects) visible and searchable on the web;
- The alignment of the data relative to persons as the first step in the process.

[https://www.transition-bibliographique.fr/fne/french-national-entities-file/](https://www.transition-bibliographique.fr/fne/french-national-entities-file/)
2. Assessment of existent authority records: poor record for certain types of authorities

The creation of the unique National Entity File is supposed to be based on the use of semi-automated alignment methods and tools.

3 problems to be expected:
- Alignment tools need to be improved taking into account the specificity of the multi-script data;
- The data are not « clean »;
- A large set of authority records (notably those including original scripts) are complicated. A particular attention should be paid to the authority records including Arabic characters (authors writing in Arabic, Persian, Ottoman Turkish, Kurdish, Urdu, etc.)

Typical anomalies with authority records in Arabic characters:
- Duplicate records due to the lack of language skills in non-specialized libraries or due to the incompleteness of the existent records;
- Duplicate records due to diverging practices (the same author in Arabic or Turkish-cataloguing context; different transliteration);
- Duplicate records for the persons changing their cultural environment during their life-time;
- Linguistic problems: spelling and vocalisation variations in Arabic script languages;
- Bad links between bibliographic and authority records;
- Lacking name authority records due to the homonymy;
- Lack of normalisation of Arabic characters in French cataloguing systems (alif maqsura and Persian ye; alif-hamza and alif-wasla)
<table>
<thead>
<tr>
<th>Lot</th>
<th>Nombre de résultats</th>
<th>Notice</th>
<th>PPN</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>1</td>
<td>1</td>
<td>227507991</td>
<td>UNM</td>
</tr>
</tbody>
</table>

**Ibn Fāris (941-1004):** duplicate record due to the lack of language skills in the cataloguing library.
Kınalızade Ali Çelebi (1510-1572) : duplicate record due to diverging practices

**Notice de type Personne**

**Point d'accès autorisé**

Kınalızade, Âli (1510-1572; Çelebi)

**Variantes de point d'accès**

Ali Çelebi, Kınalıça
don't @

**Information**

Langue d'expression: Türk ottoman (écriture arabe), Arab
Pays: Turquie, République arabe Syrienne, Égypte
Date de naissance: 1510
Date de décès: 1572
Genre: Masculin

**Notes**

Note publique d'information: Cadî de Bursa, Edime, Le Caire, Damas, Aleppo, Istanbul; paléographe ottoman; Kazasker d'Anatolie en 1571.

---

**Multilingual person names and Arabic script in authority files and research databases: the Mistara Project / Olga Andriyanova (BIMA), Asyeh Ghaforian (Bulac) & Benjamin Guichard (Bulac)**

41st IELCom International Conference, Naples, 18-20 June 2019
Siyah Kalem (lived 15th cent.), name of convenience given to a not-identified painter:
duplicate authority record due to diverging romanisation methods and the choice of the heading

Siyah Kalem, Muhammed (https://www.idref.fr/193377705)

Ustad Mehmed Siyah Qalem (actif 14..-14..) (https://www.idref.fr/160638143)
Authority record for Ibn Rushd al-Jadd, Abu al-Walid Muhammad ibn Ahmad (d. 1126) : a classical Arab Islamic author record lacking forms in Arabic characters and in ISO romanisation standard

https://www.idref.fr/159826829
Averroès (1126-1198) : bad links with bibliographic records (https://www.idref.fr/02825886X)
Ibn Rušd, ʻAbd ʻAllāh ibn Muḥammad (11-12): bad links with bibliographic records ([https://www.idref.fr/137969449](https://www.idref.fr/137969449))
3. Working together, with the right tools and with the right team, for a better future: the Mistara project partners and their potential contributions
Bring together diverse expertise and competence

- Dominican Institute for Oriental Studies (Cairo): [https://www.ideo-cairo.org/fr/](https://www.ideo-cairo.org/fr/)
- Arabic section of the Institut de recherche et d'histoire des textes du CNRS (Paris): [https://www.irht.cnrs.fr/?q=en](https://www.irht.cnrs.fr/?q=en)
- Unified documentation service (SCD) of Aix-Marseille University: [https://scd.univ-amu.fr/](https://scd.univ-amu.fr/)

With the support from the national bibliographic agencies:

- Bibliothèque nationale de France: [https://www.bnf.fr/fr](https://www.bnf.fr/fr)

Under the supervision by Bulac as the project carrier: [https://www.bulac.fr/](https://www.bulac.fr/)
Paprika and Qualinka: online authority control and verification service

- Helps to create / modify a link with the authority record in two clicks;
- Changes made are immediately visible in the IdRef authority file and the Sudoc

https://paprika.idref.fr/

- Expertise and contribution in the data normalisation process;
- Assistance for the use of semi-automated alignment tools and in data processing
Bibliostatus: URI alignment strategy for the bibliographic transition

- allows semi-automatic alignments of the bibliographic and authority data from local catalogues with the data managed by ABES and BnF

https://www.transition-bibliographique.fr/systemes-et-donnees/bibliostatus/
Search results for Ibn Rushd in data.bnf.fr

Auteurs

→ Averroès (1126-1198)
  Muhammad ibn Ahmad ibn Muhammad ibn Ahmad ibn Rušd (1126-1198)
  Savant, théologien et philosophe

→ Averroès (1058-1126)
  Muhammad ibn Ahmad ibn Ahmad ibn Rušd al-Qurtubi (1058-1126)
  Juge, spécialiste du droit islamique

Œuvre

→ Damīnāt al-Ilm al-ilāhī
  Die Frage welche Ibn Rušd in der ersten Abhandlungen berührt hat
  Averroès (1126-1198)
  Supplément au traité "Fatā al-maṣāḥīh"
BIMA and the A'LÂM name authority file

Standard authority record for a classical author from the authority file of the Arab World Institute's Library (previously A'LÂM authority file):

Ibn Rushd, Muhammad ibn Ahmad (1126-1198)

About A'LÂM authority file:

- created in 1994 as a database including a double script authority file and a bio-bibliographical dictionary in order to respond to the needs of the Arab World Institute Library;
- 14 425 records (in 2009):
  1 930 authors from Antiquity or Classical period;
  10 083 contemporary Arab authors (end 19th – 20th c.);
  223 non Arab et non European authors (India, Turkey, Iran…);
- the authority file migrated into Virtua LIS in 2013: the authority file is searchable through the OPAC (http://bima-catalogue.imarabe.org/), but the authority records themselves can’t be viewed at the moment.
IMA authorities today:

- since 2015, alignment process with IdRef authorities (PPN number introduced into the local record; enriching IdRef authorities with name variants from local records; creating new authority records in WINIbw / IdRef (more than 3000 new authority records created);
- discussions about the further action: to substitute A‘LÂM authorities with IdRef records or to align them preserving the data locally created;
- BIMA and Mistara project: continuing to enrich the IdRef authority file; providing extractions from A‘LÂM; contribution to the discussion on normalization process.
MMSH library
(Médiathèque de la Maison méditerranéenne des sciences de l’homme)

Joined Sudoc at the beginning of 2019
as an integral part of the Documentation Service of Aix-Marseille University

http://mediabase.mmsh.univ-aix.fr/flora_app/

https://catalogue.univ-amu.fr/
(contributor: Aix – Jas de Bouffan MMSH)
MMSH Library: more than 160,000 monographs, including 50,000 in Arabic, and 5,000 periodicals; http://www.mmsh.univ-aix.fr/mediatheque/Pages/default.aspx

- Search by author in the MMSH catalogue;

- MMSH’s contribution to the project: local data extraction for the Mistara project purposes; participation in the data alignment.
Al-Kindi catalogue of the Dominican Institute for Oriental Studies Library

https://alkindi.ideo-cairo.org/authority/41843
Onomasticon Arabicum :
a long-living onomastic database project
https://onomasticon.irht.cnrs.fr/

➢ Before 1914 : 250 000 onomastic files gathered in Rome
(Accademia dei Lincei, Fondazione Caetani) by Giuseppe
Gabrieli and Leone Caetani. 2 volumes published in 1915.

➢ 1960s : beginnings of computerizing on punched cards at the
IRHT

➢ 1964 : Jacqueline Sublet discovers the existence of the Italian
onomastic archive. A copy is made for the IRHT.

➢ 1966 : Announcement of the creation of a project to renew the
Italian Onomasticon Arabicum and computerizing Arabic
prosopographical data.

➢ 2002 : delivery of the onomastic database for classical authors
on Access

➢ 2010 : Onomasticon Arabicum goes online (SQL) with the
data on more than 27000 scholars and celebrities from the
first Muslim millenary


http://dar.bibalex.org/webpages/mainpage.jsf?PID=DAF-Job:145036&q=onomasticon
Onomasticon Arabicum (OA) is a long-living database project. This new online-version informs on more than 27000 scholars and celebrities from the first Muslim millenium. Its entries in Arabic are compiled from ancient biographical dictionaries, a veritable treasure of Islamic culture. Crossed search allows separate interrogation on any of the different elements of the Arabo-Muslim names, dates and places, reconstructing the identity of a person, trace ways of knowledge transmission and frame historical contexts.
<table>
<thead>
<tr>
<th>Comment</th>
<th>Content</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISM</td>
<td>مُحَمَّد</td>
<td></td>
</tr>
<tr>
<td>IAB</td>
<td>أحمد</td>
<td></td>
</tr>
<tr>
<td>GAD</td>
<td>العلامة المفتى أبو الوليد</td>
<td></td>
</tr>
<tr>
<td>ABG</td>
<td>أحمد</td>
<td></td>
</tr>
<tr>
<td>GAG</td>
<td>زُنّد</td>
<td></td>
</tr>
<tr>
<td>KUN</td>
<td>أبو الوليد</td>
<td></td>
</tr>
<tr>
<td>LQB</td>
<td>العلامة</td>
<td></td>
</tr>
<tr>
<td>SUH</td>
<td>ابن زُنّد الخفيد</td>
<td></td>
</tr>
<tr>
<td>DIN</td>
<td>المالكي</td>
<td></td>
</tr>
<tr>
<td>NSB</td>
<td>الفرطري</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>طبيب</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>تقية</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>أقبل على الفلسفة</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>فِيْلَسُوف</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>أقبل على الكلام</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>مَنَّكَمُ</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>سَمِعَ الحديث</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>مُحَدِّث</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>في الأثاثي</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>مَصَنِّف</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>في الفقه والطَّبّ</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>مَصَنِّف</td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>520</td>
<td>WLD</td>
</tr>
<tr>
<td>MAT</td>
<td>595</td>
<td></td>
</tr>
<tr>
<td>REF</td>
<td>SD, t. 6, p. 522</td>
<td></td>
</tr>
<tr>
<td>REF</td>
<td>SD1, t. 4, p. 320</td>
<td></td>
</tr>
</tbody>
</table>
General

ISM First name (ism ʿalam)
IAB First name of parent (ism al-ab)
GAD First name of grandparent (ism al-ḡadd)
ABG First name of great-grandparent (ism ab al-ḡadd)
GAG First name of great-great-grandparent (ism ab al-ḡadd)
GNS Contains أرمة: feminine gender
GGG Shows there is more ancestors’ names in the text
KUN Name elements with abū or umm (kunya)
LAQ Honorific laqab containing al-Dīn
LQB Other honorifics, titles, nicknames (laqab, talqīb, naʿt)
SUH Name by which individual is commonly known (šuhra)
DIN Religion, spiritual order, legal school (dīn, tariqa, maḏhab)
NSB Name elements ending in ī (nisba)

Places

HAL Places of residence (maḥall al-sakn)
HDF Place of burial (maḥall al-dafn)
HDR Places of teaching (madāris)
HRB Places of military activity (maydān al-ḥarb)
HRI Places visited (riḥlāt)
HWL Place of birth (maḥall al-wilāda)
HAQ Places or regions of ascendance or descendance (maḥall al-ʿaqb)

Dates

WLD Date of birth (taʿrīḫ al-wilāda)
MAT Date of death (taʿrīḫ al-wafāt)
TRH Date of event, other than birth or death (taʿrīḫ)

More

AHL Family or tribal relations (ahl)
AKH Moral/religious qualities (aḥlāq)
AQB Information on ancestors or descendance (ʿaqb)
EGL Shows that two or more notices concern the same individual
ILA Teacher, transmitter (rawā ilā)
MIN Student, transmitter (samiʿa min)
REF Bibliographical reference
SIF Physical attributes (ṣifa)
SLT Relationships between persons (ṣila)
SMT Cause of death (asbāb al-wafāt)
SNF Title of book composed by the individual (taṣnīf)
SRK Hobbies (ištirāk)
SWMOccupation (ṣuğl, waṣīfā, manṣīb)
UMT Age at death (ʿumr)
XXX Shows all the rubrics except REF (used only for a search)
Mistara project: work together to improve the IdRef authority file

- Define a set of records for intervention (names of persons in Arabic script registered in several linguistic environments, but not only);
- Clean, deduplicate, standardize and enrich the existent authority data, combining semi-automated and manual methods;
- Create new authorities on the basis of the existent onomastic databases (Al-Kindi, Onomasticon);
- Attribution of ISNI (International Standard Name Identifier);
- Prepare the alignment procedures with other databases.
4. Desire to ensure a quality service for the research community and for the library professionals

➢ Ensure the evolution of the alignment tools for bibliographic transition and their taking into account of the linguistic complexity of multilingual person names in Arabic script ;
➢ Propose a linchpin data schema between IDref, Alkindi and Onomasticon Arabicum ;
➢ Enrich IdRef (and thus the future FNE) to make it a high-quality resource for the researchers ;
➢ Link together the onomastic research and the ‘cleaning’ of the authority reference files
➢ Enrich the RDA-FR good practice guidance for identifying the preferred name headings, taking into account the existent standards (e.g. *IFLA Names of Persons* : [https://www.ifla.org/node/4953](https://www.ifla.org/node/4953) ; *Arabic NACO manual* : [http://library.princeton.edu/departments/tsd/katmandu/cp20/aranatoc.html](http://library.princeton.edu/departments/tsd/katmandu/cp20/aranatoc.html)) and detailing them on the basis of the specific cases studied in the frame of the project.
Mistara project today: we are at the very beginning

- Financing obtained through Collex-Persée (Collections of excellence for scientific Research: http://www.collex.eu/)

- 11 July 2019 – first meeting of the project’s steering committee

- Project duration – 18 months (2019-2020)

Challenges to meet:

- ensure the quality of the data and the metadata;
- Create a dialogue between the structured data and their enrichment;
- Push the boundaries created by the alignment tools and the manual data treatment.
Thank you very much for your attention!

mistara@bulac.fr