JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12727

DREAM. A project about non-Latin script data

Antonella Fallerini[a], Agnese Galeffi[b], Andrea Ribichini[c], Mario Santanché[d], Mattia Vallania[e]

a) Sapienza Università di Roma
b) Sapienza Università di Roma, https://orcid.org/0000-0003-0799-5699
c) Sapienza Università di Roma, https://orcid.org/0000-0002-0281-4257
d) Sapienza Università di Roma, https://orcid.org/0000-0003-1777-1162
e) Sapienza Università di Roma

Contact: Antonella Fallerini, antonella.fallerini@uniroma1.it; Agnese Galeffi, agnese.galeffi@uniroma1.it; Andrea Ribichini, ribichini@diag.uniroma1.it; Mario Santanché, mario.santanche@uniroma1.it; Mattia Vallania, mattia.vallania@uniroma1.it

Received: 5 May 2021; **Accepted:** 16 June 2021; **First Published:** 15 January 2022

ABSTRACT

The DREAM project is a large research project funded by Sapienza University of Rome, dealing with bibliographic data in non-Latin scripts. As the National Bibliographic Service catalogue (SBN) does not yet manage data in non-Latin scripts, the aim of DREAM is to offer researchers a catalogue searchable through original scripts (such as Arabic, Chinese, Cyrillic, etc.). One of the most remarkable features of the project is the creation of an ILS-independent working context in which the cataloguer may find and retrieve data in original script from authoritative catalogues, starting from the existing romanized ones. From a technical standpoint, the ever increasing Unicode support offered by modern operating systems, DBMSs and indexing engines makes the rapid development of the relevant software tools a concrete possibility. This in turn implies a shift in scientific focus towards the (often subtle) record linkage operations between different data sources. The authors hope that the DREAM project will gather the adhesion of other Italian libraries that perceive the same needs. Furthermore, as soon as SBN will support the management of data in non-Latin scripts, the DREAM project partners will be able to contribute with their data.

KEYWORDS

Romanization; MARC records; Cataloguing; Transliteration.

Non-Latin script cataloguing. The context

The DREAM (Data Recording Entry Alternative Multi-script) project was born in Sapienza university in 2019 in order to create a repository for bibliographic data in non-Latin scripts, publicly available as a cooperative catalogue. This need arises from the evidence that the SBN national catalogue, to which Sapienza libraries adhere as do thousands of other Italian libraries, does not fully support the UTF-8 character encoding. SBN catalogue is based on shared cataloguing: all the participant libraries contribute sending data from the local nodes (that is, aggregation of libraries) to the central index. Member libraries may use a variety of LMS authorized by the ICCU, the national agency in charge with the SBN catalogue maintenance, but regardless of the software capabilities, the central index accepts data in Latin script only. All the languages expressed through other scripts, such as Arabic, Chinese, Cyrillic, Hebrew, Japanese, Greek, must be transliterated using the ISO instructions. This requirement is stated in the Italian cataloguing rules, REICAT (ICCU 2016b), and restated by the SBN cataloguing instructions (ICCU 2016a). The transliteration process has many disadvantages for both the involved actors – cataloguers and users.

Notwithstanding some attempts at automatic transliteration (Eryani 2021) and the availability of online tools (DuBose 2019), this activity is a very time consuming one presenting a large number of technical problems (Ismail and Md. Roni 2010), not to mention the variety of connected cataloguing issues such as

- The use, in some contexts, of unsound transliteration scheme (Molani 2006).
- The transliteration and conversion of personal names, place names, corporate bodies, and other entities (Li 2004).
- The subject access (El-Sherbini and Chen 2011).

Besides that, the equity of access – one of the basis of the ethics in library science – is not guaranteed since those who need these materials have to determine how cataloguers could have transliterated the data. On the opposite, the users who want roman script resources are not required to make this extra and inefficient effort (Agenbroad 2006, 22). For users who are native language speakers in non-Latin scripts, the transliteration is totally useless since they have all the knowledge and tools to perform a search in the library catalogues using the original script.

For all other users reading in second language, transliterated text requires an additional cognitive demand since they typically acquire and access a cohesive set of phonological, orthographic (and possibly semantic) representations of words in their second language, whereas transliteration requires readers to create cross-script associations between phonological-semantic representations in one language and previously unrelated orthographic forms in another (Rao, Mathur, and Singh 2013, 205). All these elements cause many obstacles in users' searching ability and accessibility in retrieving bibliographic records (Kim 2006)

The DREAM project

The DREAM project is a major university project funded by Sapienza University of Rome in 2019; the project leader is Federico Masini, professor of Chinese at the Istituto Italiano di studi orientali (ISO) of Sapienza university of Rome and the research staff is a mix of academics and library personnel, both involved on the front line of the activities. The very first idea of the DREAM project arose in May 2018, when Antonella Fallerini, librarian at the ISO library, joined the workshop "Building a Network of Korean Resources Specialists in Europe", organized by Freie Universität Berlin – Campus Library and funded by the Korea Foundation. The workshop aimed at bringing together European Korean Studies librarians in order to develop a professional network within Europe and strengthen the representation of interests of Korean Studies librarians in national and worldwide library information structures. While discussing with the colleagues attending the workshop, Fallerini highlighted the severe limitations of transliterated data compared to bibliographic descriptions in original scripts. The expected result of that was that some colleagues confirmed they did never find a single record in original script in Italian catalogues. Their obvious self-explanation was that there was a great scarcity of our collections in Far Eastern languages. The available online catalogues do not give any advice about the transliteration and researchers have no reason to expect such a treatment. The extensive transliteration practice in cataloguing has as a direct consequence the underrepresentation of our library collections both at national and international level. To give an idea of the extend of the phenomenon, an internal preliminary investigation conducted on the online catalogues of the most representative Italian institutions, such as larger universities and research libraries, have shown that more than 500,000 resources have been catalogued in romanization. It is possible to estimate that there are at least double that number waiting to be catalogued.

What is the aim of DREAM?

DREAM project aims to figure out a provisional and cooperative solution in order to create a repository for non-Latin scripts data, available as a catalogue in the near future. At the present moment, the cataloguers must create transliterated data to feed SBN; if adhering to the DREAM project, the cataloguer will also use DREAM tools to search for the corresponding record in original script in authoritative catalogues. Once the possible matches have been identified, the system will present them to the cataloguer, giving him/her the responsibility of confirming of dismissing them. These data will complement the transliterated ones that are already being produced for SBN shared cataloguing.

The result of this procedure will be a cluster of records that will be show in the DREAM catalogue giving the user the possibility to make searches using the preferred forms (in original scripts or in transliteration, even according to different schemas).

The DREAM project do not want to propose an SBN-alternative context. On the opposite, its features are developed taking into consideration both the respect of SBN rules and its potential developments. When SBN will accept data in non-Latin scripts, the libraries adhering to DREAM will have the possibility to feed their records into the national catalogue. This is why the DREAM project has a provisional nature. The adjective "provisional" may be referred to two aspects: first of all, it connotes the research aspect. DREAM's aim is to produce a working solution and at the same time, to explore, to verify, and to find the best ways to achieve the projects' goals. Since Sapienza libraries are part of the SBN network, there is no intention to create a new network or some alternative solutions; this is the second significance of provisional. DREAM project wants to create an environment where the cataloguer can retrieve data in non-Latin scripts, match them with the available transliterated ones and make them available in a specific DREAM catalogue. Both the working environment and the user search interface will be independent from SBN as well as from the software used by the libraries joining the project. We hope in fact that other Italian libraries – even those not members of SBN – will be interested in joining the DREAM catalogue, once some of the fundamental components of the its architecture have been realized. The DREAM project is still ongoing. We would like to stress one of the project's strengths: flexibility. We have a clear idea of the final results we want to achieve, but there is no prejudice about how to reach them. There are just some constrains due to the cataloguing context we have to dialogue with at some point, that is the software used and the SBN catalogue.

Main points

DREAM is ILS independent

Commercial software available to librarians are built to maximise output (cataloguing, lending, library management, etc.) and are therefore, in most cases, designed to be stable and standard. If you need a flexible environment to, for instance, carry out an experiment or a research project, it is difficult to balance these development needs – maybe even unsuccessfully – with the commercial logic of software distributors. Anyway, Sapienza has invested in our ILS (SebinaNext) in order to implement in the near future some new features, such as to accept, manage and visualize data in non-Latin script and especially right-to-left scripts, to handle VIAF id and an OAI-PMH module for authority data. DREAM will be an external and ILS independent environment. This need would not have arisen if we had a flexible and welcoming open source software or library platform in use. In this case, the DREAM project would have been just another component, a small one, of a larger system. What we learnt: often the paths you thought you were taking do not turn out to be fruitful and you have to go back, change your path and sometimes even rewrite the map. These features of research projects do not match the market logic.

Retrieve bibliographic data from reliable sources

In order to quickly populate the DREAM catalogue, we are going to start from the traditional transliterated records already existing, to search for equivalent records in original script in authoritative catalogues, and to import them. These procedures present certain degrees of difficulty. First of all, the identification of reliable sources to retrieve data. This is a scientific but also a technical task. It is not only a matter of knowing the most representative institutions for the languages of interest, but also of selecting those that have a data format easy to manage or map and an accessible retrieval option.

The current DREAM implementation supports this "search and match" between, on one side Sapienza University of Rome catalogue and on the other side the Bibliothèque nationale de France, the Library Union Catalogue of Bavaria, Berlin and Brandenburg (B3kat), and the Système universitaire de documentation (SUDOC). Since these sources expose their data through a variety of protocols, such as OAI-PMH, SRU and Z39.50, different clients are needed. Moreover, to process data, specific response parsers are required for each source. As a matter of fact, even though the retrieved data are in standard formats (MARC21, UNIMARC), the packaging of data varies from source to source, containing error messages and paging information in ad hoc formats. Even in environments that we assume to be highly standardized (dealing with MARC, Z39.50, SRU, OAI-PMH formats) we found, in addition to the expected MARC21-UNIMARC dichotomy, USMARC or local dialects of MARC, Dublin Core, and several application profiles. In order to obtain a presumed match of the data, different analyses and mappings are required each time for their retrieval and processing.

Different sources (we are talking about national bibliographies/catalogues and national library catalogues) also have different approaches to standards.

For example, MARC21 allows to put in the same record data in the original script (e.g. Cyrillic) together with transliterated data by using the combination 880 and $6 but the cataloguing agency can choose whether to put in 880 the original script or the transliterated version. This allows the creation of (at least) two versions of the record. Moreover, the different granularity of the data contributes to make the match uncertain.

Authority data

Obviously, within the DREAM environment, in addition to bibliographic data, it is essential to import, manage and use authority data. In this respect, VIAF is the point of reference. Since the VIAF id is widely used, it is not only possible to retrieve authority clusters, but also to use the VIAF id as a bridge to navigate through catalogues in search of other bibliographic data of potential interest.

What we are building. The DREAM architecture

We designed a flexible, modular and scalable software architecture for a multiscript MetaOPAC (see Figure 1), based on the data warehousing paradigm (Inmon 2005; Kimball et al. 2008). We also developed a prototype implementation for research purposes (i.e., feasibility assessment, experimental evaluation of adopted solutions). The following is the description of our architecture's main components.

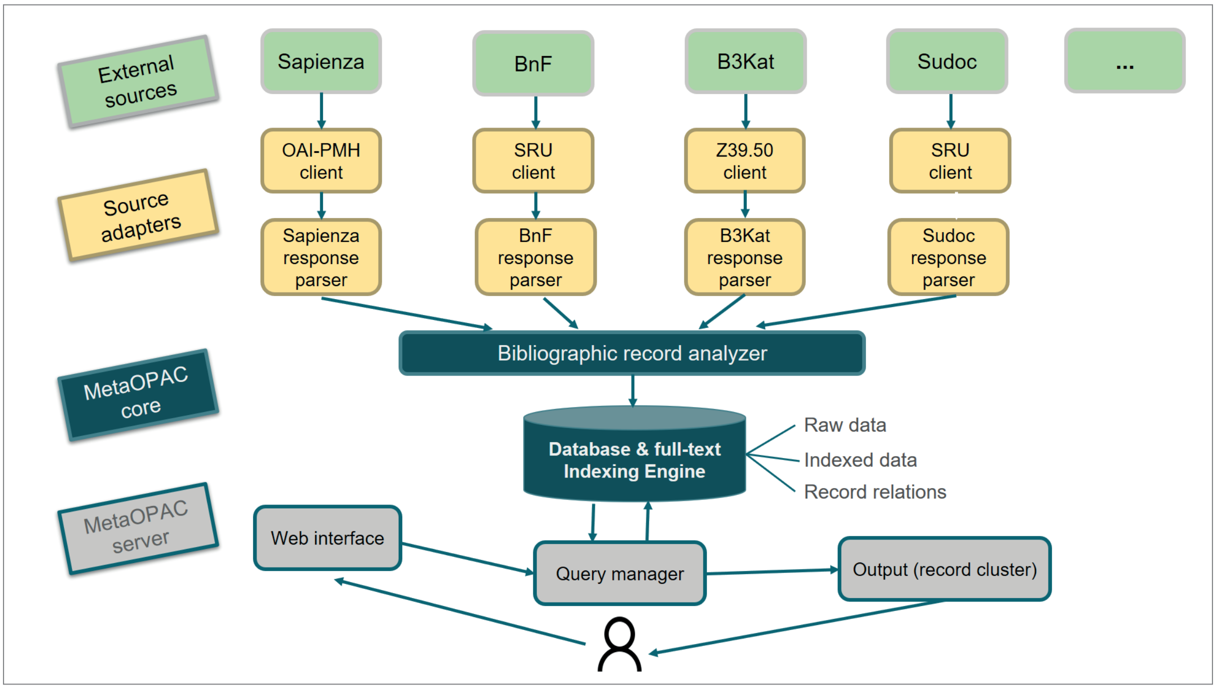

Fig. 1. MetaOPAC architecture

Source Adapters. We have taken into consideration and tested several data sources. The current implementation supports Sapienza University of Rome's own catalogue, the Bibliothèque nationale de France (BNF), the Library Union Catalogue of Bavaria, Berlin and Brandenburg (B3Kat), and the Système universitaire de documentation (SUDOC). These sources make their data available through a variety of protocols, such as OAI-PMH, SRU and Z39.50. Therefore, clients are needed for each of these protocols. Moreover, an ad hoc response parser is required for each data source. This is because, even though the returned data are provided in standardized formats (e.g., UNIMARC, MARC 21), the packaging of these formats varies from source to source.

MetaOPAC Core. Downloaded bibliographic records are stored in a (relational) database, analyzed, and portions that are relevant for future search queries, e.g., title, authors, publisher (including all variants in both native script and transliteration, if present) are saved separately and properly indexed. Our prototype implementation currently uses MySQL as DBMS. The database structure consists of three tables:

- Table "raw" contains the unprocessed downloaded records.
- Table "indexed_data" contains, for each record, the extracted data to be indexed in order to speed up searches. At the present moment, we rely on MySQL's full-text indexing capabilities (a recent addition). We remark that different scripts require different indexing methods: alphabetic and syllabic scripts are handled by the default token-based full-text indexer, with minimum token size set to 1 and stop words exclusion disabled, while Ideographic scripts are instead dealt with by an n-gram based indexer, with n=2.
- The third database table, "relations" represents associations between records from different data sources, that we call "clusters". Clusters may be established through several methods (that we will discuss shortly).

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12727

MetaOPAC Server. Searches in our prototypal MetaOPAC implementation can be run through a web server that accepts HTTP GET requests. In addition to the traditional search criteria (keywords, title, author, publisher), wildcards and boolean operators are accepted. A query manager translates the searches into full-text database queries. The search results are returned as an XML document listing retrieved clusters sorted by *relevance* (a measure of the adherence of the records in each cluster to the search criteria).

How to feed the DREAM. Record linkage among data sources

In our MetaOPAC application, the construction of clusters (i.e., groups of records referring to the same entity) may be carried out through three methods.

1. *Manual Intervention.* The cataloguer manually identifies the correspondences between records from different data sources. In our prototype we have created 27 Sapienza-BNF pairs, 27 Sapienza-B3KAT pairs, and 40 Sapienza-SUDOC pairs. It is hoped that, as the number of partners grows, more and more librarians will contribute their associations across data sources to the MetaOPAC database.
2. *Identification by Unique Identifiers.* A second way to identify correspondences between records from different data sources is through unique identifiers. In our prototype we have used ISBN to search for matches (all supported external catalogues allow ISBN-based searches through their APIs).

Document Language	Sapienza Records with ISBN	Sapienza-BNF ISBN-based Matches	Sapienza-B3Kat ISBN-based Matches	Sapienza-SUDOC ISBN-based Matches
ARA	369	122 (33.06%)	113 (30.62%)	126 (34.15%)
CHI	1875	98 (5.23%)	492 (26.24%)	399 (21.28%)
HIN	25	8 (32%)	3 (12%)	8 (32%)
JPN	1771	246 (13.89%)	692 (39.07%)	781 (44.10%)
KOR	2191	80 (3.65%)	432 (19.72%)	457 (20.86%)
PER	66	7 (10.61%)	12 (18.18%)	15 (22.73%)
SAN	73	17 (23.29%)	26 (35.62%)	28 (38.36%)
SWA	1	1 (100%)	1 (100%)	1 (100%)

Table 1. Breakdown of positive search results

3. *Algorithmic Techniques.* The third method consists of a blend of well-established record linkage algorithmic techniques and ad hoc solutions. We proposed the following workflow, based on the VIAF:
   - Given as input a bibliographic record, we extract the VIAF code of its author (assumed to be present).
   - We then run a search on the VIAF online service for the extracted id, obtaining the variant form of the author's name used by each data source.

- For each supported source, a search-by-author, using the variant form obtained through VIAF as input string, is performed. This allows us to restrict the search domain to the works of that author.
- Finally, we run any record linkage algorithm we see fit in order to identify the correct matches between the input record and the records retrieved from the other data sources.

Standard record linkage techniques include the use of string similarity measures (Navarro, 2001) – Levenshtein distance (Levenshtein, 1966) being a popular one – to assess correspondences between fields such as title, subtitle and publisher (including their variants and versions in original script, if present). Comparison of other metadata (e.g., publication dates) may also be useful as a verification tool. Moreover, if the bibliographic record belongs to a cluster in the MetaOPAC database, then all metadata of the cluster may be used to identify the correct match. More sophisticated, domain-specific techniques may include transformations from one transliteration standard to another, and switching from original script to transliteration and vice versa. Early testing on 19 Sapienza records, manually matched with both BNF and SUDOC to provide a "ground truth", has shown correct results in 17 cases. This is quite promising considering that for this test only the minimum normalized Levenshtein distance (i.e., Levenshtein distance divided by the length of the longest input string) between all title variants has been considered as a criterion.

Further steps

The project next steps are:

- Engaging partner institutions: we hope that this conference will also be an opportunity to promote the project and involve other partners who share the problem with data in non-Latin scripts
- From a technical standpoint, further tasks would include writing adapters to support additional sources, and launching larger scale algorithmic record linkage runs with feedback loops involving manual sample validation and fine-tuning of algorithmic features. Identified clusters should then be fed into the MetaOPAC prototype implementations, with measurement of both load and query times, in order to determine performance-critical sections that may need refinement both at the implementational and the architectural level.
- It is also needed to develop all the interfaces, both the back office minimal interface to allow cataloguers to validate the matches between records and the public DREAM catalogue search interface.

Bibliography

Agenbroad, James E. 2006. "Romanization Is Not Enough." *Cataloging & Classification Quarterly* 42 (2): 21-34. https://doi.org/10.1300/J104v42n02_03

DuBose, Joy. 2019. "Russian, Japanese, and Latin Oh My! Using Technology to Catalog Non-English Language Titles." *Cataloging & Classification Quarterly* 57 (7-8): 496-506. https://doi.org/10.1080/01639374.2019.1671929

El-Sherbini, Magda, and Sherab Chen. 2011. "An Assessment of the Need to Provide Non-Roman Subject Access to the Library Online Catalog." *Cataloging & Classification Quarterly* 49 (6): 457-483. https://doi.org/10.1080/01639374.2011.603108

Eryani, Fadhl, and Nizar Habash. 2021. "Automatic Romanization of Arabic Bibliographic Records." https://arxiv.org/pdf/2103.07199.pdf

ICCU. 2016a. "Guida alla catalogazione in SBN – Materiale moderno." Last modified July 13, 2016. https://norme.iccu.sbn.it/index.php?title=Guida_moderno/Descrizione/Capitolo_generale/Lingua_e_scrittura_della_descrizione

ICCU. 2016b. "Regole italiane di catalogazione. Appendice F – Traslitterazione o trascrizione di scritture diverse dall'alfabeto latino." Last modified September 21, 2016. https://norme.iccu.sbn.it/index.php?title=Reicat/Appendici/Appendice_F

Inmon, William H. 2005. *Building the data warehouse*. 4th ed. Indianapolis: John Wiley & Sons.

Ismail, Mohd Ikhwan, and Nurul Azurah Md. Roni. 2010. "Issues and challenges in cataloguing Arabic books in Malaysia academic libraries." *Education for Information* 28 (2-4): 151-163.

Kim, SungKyung. 2006. "Romanization in Cataloging of Korean Materials." *Cataloging & Classification Quarterly* 43 (2): 53-76. https://doi.org/10.1300/J104v43n02_05

Kimball, Ralph, Margy Ross, Warren Thornthwaite, Joy Mundy, and Bob Becker. 2008. *The data warehouse lifecycle toolkit*. 2° ed. Indianapolis: John Wiley & Sons.

Kudo, Yoko. 2010. "A Study of Romanization Practice for Japanese Language Titles in OCLC WorldCat Records." *Cataloging & Classification Quarterly* 48 (4): 279-302. https://doi.org/10.1080/01639370903338352

Levenshtein, Vladimir Iosifovich. 1966. "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady* 10 (8): 707-710.

Li, Yue. 2004. "Consistency versus Inconsistency: Issues in Chinese Cataloging in OCLC." *Cataloging & Classification Quarterly* 38 (2): 17-31. https://doi.org/10.1300/J104v38n02_04

Molavi, Fereshteh. 2006. "Main Issues in Cataloging Persian Language Materials in North America." *Cataloging & Classification Quarterly* 43 (2): 77-82. https://doi.org/10.1300/J104v43n02_06

Navarro, Gonzalo. 2001. "A guided tour to approximate string matching." *ACM Computing Surveys* 33 (1): 31-88. https://doi.org/10.1145/375360.375365

Rao, Chaitra, Avantika Mathur, and Nandini C. Singh. 2013. "'Cost in Transliteration': The neurocognitive processing of Romanized writing." *Brain and Language* 124 (3): 205-212. https://doi.org/10.1016/j.bandl.2012.12.004