DREAM. A project about non-Latin script data

A. Fallerini, A. Galeffi, A. Ribichini, M. Santanché, M. Vallania
(Sapienza Università di Roma)

Abstract

The DREAM project (https://web.uniroma1.it/sbs/en/dream) is a large research project founded 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.

What is DREAM? Project overview

The DREAM (Data Recording Entry Alternative Multi-script) project was born in Sapienza University of Rome a couple of years ago. It has been funded as a major university project and this constitutes a brand new feature: it is the first Sapienza major project in which library staff are involved in the front line of research together with academic staff.

The idea of DREAM was born in May 2018, when our colleague Fallerini 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. In that occasion, discussions with other librarians highlighted the severe limitations of transliterated data compared to bibliographic descriptions in original scripts. Some colleagues even said they could never find a single record in Italian catalogues and wondered why, marveling at the scarcity of our collections in Far Eastern languages.

A survey conducted on the online catalogues of the most representative institutions, such as larger Italian universities and other cultural institutions, shows that more than 500,000 resources have been catalogued in romanization. We have also estimated that there are at least double the number of resources waiting to be catalogued.

What is the aim of DREAM?

It is all about UTF-8, a character encoding widely used, but unfortunately not fully supported by SBN national catalogue, to which Sapienza libraries adhere (as do thousands of other Italian libraries). SBN is based on shared cataloguing, but member libraries are free to use a variety of ILS software to send/retrieve records to and from the centralized database. At the present moment, the SBN catalogue does not accept data in non-Latin scripts (such as Arabic, Chinese, Japanese, Korean, Cyrillic, etc), therefore cataloguers are obliged to transliterate them. As you can easily imagine, this activity is time consuming for the cataloguer, but it is also not useful for the researcher. Moreover, the transliterated data have limited international relevance as researchers all over the world do not expect to have to use the transliterated forms when searching for library materials.

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. These data will complement the transliterated data that are already being produced for SBN shared cataloguing. When SBN will be able to accept data in non-Latin scripts, the libraries adhering to DREAM will have the possibility to feed their data into the national catalogue.

We have said that DREAM is intended to be a project of a provisional and cooperative nature. What does it mean? "Provisional" has two meanings.

First of all, provisional connotes the research aspect. DREAM "is" a research project. Obviously the aim is to produce something that works but at the same time, to explore, to verify, and to find the best solutions to achieve our goals.

The second meaning of "provisional" is that Sapienza libraries are part of the SBN network and there is no intention to create a new network or some alternative solution. DREAM project wants to create an environment where the cataloguer can retrieve data in non-Latin scripts and make them available in a specific catalogue. This working environment and the user search interface will be independent from SBN as well as from the software used by the libraries that want to participate in the project. We hope in fact that other Italian libraries - even those not members of SBN - will be interested in joining the DREAM catalogue.

“Cooperative” implies that we would like, once some of the fundamental components of the DREAM architecture have been realized, to involve other institutions in enriching the DREAM catalogue.

The DREAM project is still ongoing. I 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 we have no prejudice about how to reach them. We just have some constraints due to the cataloguing context we have to dialogue with at some point, that is the software we use and the SBN catalogue.

How to bulk-feed data in non-Latin script into a catalogue while having the equivalent transliterated data available? Obviously starting with the present data and using them to retrieve non-Latin script data from authoritative catalogues.
On the positive side, there are many procedures, tools and solutions to achieve this workflow.
On the negative side, there are many procedures, tools and solutions to achieve this workflow.

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 ambient. 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 plan, as explained later from a technical point of view, to start from the dear old transliterated records we already have, search for equivalent records in original script in authoritative catalogues and import them. Sounds easy?

First of all, the identification of reliable sources from which to search and retrieve data is crucial. 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, producing errors and paging information.

What we learned: in environments that we assume to be highly standard (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 reprocessing.

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.

- The different granularity of the data makes 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, based on the data warehousing paradigm. We also developed a prototype implementation for research purposes (i.e., feasibility assessment, experimental evaluation of adopted solutions).

Source Adapters. We have taken into account and tested (both successfully and unsuccessfully) several data sources. The current implementation supports matches from Sapienza University of Rome's own catalogue to 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:

1. Table "raw" contains the unprocessed downloaded records.
2. 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.
3. The third database table, "relations" represents associations between records, that we call "clusters". These clusters may be established through several methods (that we will discuss shortly).

MetaOPAC Server. Searches in our prototypal MetaOPAC implementation can be run through a minimal 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 anche match (luckily all external catalogues allow ISBN-based searches through API). A breakdown of positive search results is reported in the table below.

Document Language	Sebina Records with ISBN	Sebina-BNF ISBN-based Matches	Sebina-B3Kat ISBN-based Matches	Sebina-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%)

3. *Algorithmic Techniques.* The third method consists of a blend of classic record linkage algorithmic techniques and ad hoc solutions. We proposed the following workflow, based on the Virtual International Authority File (VIAF):
   - Given as input a bibliographic record, we extract the VIAF code of agents (assumed to be present).
   - We then run a search on the VIAF online service for the extracted id, obtaining the variant forms of agent's names used by each data source.
   - For each source handled, a search using the names' string obtained through VIAF is performed. This allows us to narrow and focus the search domain.
   - Then the appropriate record linkage algorithm is used in order to identify the correct matches between the input record and the records retrieved from the other data sources through the search-by-author query.

Standard record linkage techniques include the use of string similarity measures (Levenshtein distance 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 techniques include text normalization, transformation from one transliteration standard to another and switching from original script to transliteration or 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 criterion.

Further steps

At present, the DREAM project is still a research project, even if the aim is obviously to create a derivative cataloguing and matching records system and a catalogue.

Our next steps:

- 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 steps 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.
- We also need 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.