

EICSTES DELIVERABLE D4.2

[Miri@d](#) - A platform for digital databases analysis

Xavier POLANCO, Dominique BESAGI, Ivana ROCHE

CONTENT

1. Introduction to [Miri@d](#)
 2. [Miri@d](#) organization
 3. Resources of [Miri@d](#)
 - 3.1. Documents Delivery Management System & Customer Management System
 - 3.2. [Article@INIST](#) bibliographical databases
 - 3.3. Library Management System
 - 3.4. [Article@INIST](#) log data files
 4. [Miri@d](#) databases
 - 4.1. **QUE**ries database
 - 4.2. **DIS**play database
 - 4.3. **ORD**ers database
 - 4.4. **BIBL**iographical database
 - 4.5. **STAT**istical database
 5. Statistical module of [Miri@d](#)
-

1. Introduction to Miri@d

The purpose was to design and develop a computer-based informetric analysis server applying descriptive statistical techniques to analyse, in particular, the frequency and distribution of both Web user searches (webmetrics) and target bibliographic data (following traditional bibliometrics). It is called Miri@d. This server applies the principle of so-called “metaphoric” interfaces, that is, a man-machine interface that allows the analyst to work without needing complex training in procedures and commands.

Miri@d produces descriptive statistical data about, on the one hand, the Web user searching behaviour and, on the other hand, what is effectively used from the scientific and technical multi-disciplinary source of information available as a digital bibliographical database. The results will be made available to analysts, who can use this descriptive statistical information as a source for their indicator design tasks, and as input for multivariate data analysis, clustering analysis, and mapping; also to managers, to improve management and decision-making.

We call analysis the attempt to identify the useful information. That which is relevant to the user or expert, from a large quantity of collected data. The analysis is the common denominator of operations in which the information represents a raw material. We must process this raw material in order to extract the useful information. In our case, this raw material is represented by 7 million bibliographical references stored in a digital database, growing each day at the rate of around 3,000 new references, and by around 4,000 searches daily. These two items interlock by means of on-line transactions, which are stored, and we can now to analyse it.

In this schema we can identify, on the one hand, the resources from which [Miri@d](#) collects data:

- ✓ DM, the system of documents delivery management;
- ✓ CM, the system of customers management;
- ✓ LM, the system of library management and
- ✓ [Article@INIST](#) server that contributes with both
 - Bibliographic database
 - Its log-files.

And, on the other hand, the [Miri@d](#) databases obtained:

QUE, containing data related to queries;
DIS, containing data related to displayed bibliographical records;
ORD, containing data related to ordered documents;
BIBL, containing bibliographical records;
STAT, containing a priori calculated statistical indicators.

[Miri@d](#) is an Apache server under UNIX operating system (Solaris2) sending HTML pages dynamically generated by CGI Perl scripts and managing users' directories and sub directories.

3. The resources

The data from the different resources are received as text files from which the pertinent information is extracted, converted and formatted by shell scripts and/or Perl programmes using Ilib modules. Ilib is a library of programmes specially developed at INIST.

3.1. Documents delivery management system & customer management system

The data we can obtain from DM and CM systems must be treated in order to match information about customers document orders and customers characteristics like their geographical location or activity. After adequate treatments we have, for each document ordered, a record in a [Miri@d](#) database. This database is named ORD in the schema. Actually, these treatments have an annual periodicity and are applied, in year N, to the total number of orders received during year N-1.

3.2. [Article@INIST](#) bibliographical databases

The bibliographic data coming from [Article@INIST](#) databases are obtained by extraction under request and activated by the identification of each either new document ordered or new bibliographic notice displayed.

[Article@INIST](#) allows users to search for documents (journal articles, journal issues, journal titles and monographs) in the catalogue of the French institute for the scientific and technical information - Institut de l'Information Scientifique et Technique of the CNRS (<http://www.inist.fr> ; <http://www.cnrs.fr>). The consultation of [Article@INIST](#) is free and non-stop 24 hours on 24.

[Article@INIST](#) is structured in different sub-sets corresponding to a classical documentary description. The documents corresponding to the references constituting [Article@INIST](#) are available in INIST. The search engine employed by [Article@INIST](#) is Search97 of Verity (<http://www.verity.com>)

The users have a choice between two search modes: simple or expert. If they choose the expert one, users have four search possibilities. They can search:

- ✓ A journal title, in the database “Journals” containing journal titles;
- ✓ A particular journal issue, in the database “Issues” containing the issues of each journal;
- ✓ A determined paper, in the database “Articles” containing the articles of each issue of each journal;
- ✓ A monograph, in the database “Monographs”.

If users choose the simplified search mode, their queries run on the databases containing articles and monographs.

In the two search modes, the users can impose a limit on the publication years. The users can memorize their queries.

If a user wish to order directly a copy of documents found he must beforehand open an electronic order form account (<http://services.inist.fr/public/eng/depart.htm>). In this case, he can also use the orders management functionalities proposed by this service.

[Article@INIST](#) is updated every day (around 10,000 new records loaded by week) and it receives around 4,000 queries a day. The table 3.1 gives the content of each database.

Database	Total records (<i>December 2001</i>)
Journaux	26 497
Issues	718 960
Articles	6 086 768
Monographs	169 780

Table 3.1

Concerning the monographs we have the following distribution:

Monograph type	Total (<i>December 2001</i>)
Dissertations	92 523
Reports	41 065
Congress communications	27 307
Books	8 883

Table 3.2

These extracted bibliographic data coming from [Article@INIST](#) form a [Miri@d](#) proper bibliographic database, named BIBL in the schema. Such bibliographical records exist as XML documents:

```

<ARTICLE>
  <HEADREC>
    <ORIREC>
      <ORGSIGLE>M</ORGSIGLE>
      <DATE-INTEG datetype='6' flux='R'>17/12/1998</DATE-INTEG>
      <NOREC flux='R'>00022784</NOREC>
      <CPTNOREC flux='R'>35400007105613</CPTNOREC>
    </ORIREC>
    <DOC>
      <TYPE flux='R'>PERIODIQUE</TYPE>
      <LANGUAGE LNGCODE='fre' flux='R'>fran&ccedil;ais</LANGUAGE>
    </DOC>
    <IDDOC flux='R'>0040</IDDOC>
  </HEADREC>
  <SERFRONT>
    <SERTITLE ISDS='Y' flux='R'>Pour</SERTITLE>
    <STITLE ISDS='Y' flux='R'>Pour</STITLE>
    <SERPUBFR>
      <PUBNAME>
        <ORGNOME flux='R'>Groupe de recherche pour l'education et
          la prospective</ORGNOME>
        <CITY flux='R'>Paris</CITY>
      </PUBNAME>
      <LOCATION>
        <COUNTRY CNYCODE='FR' flux='R'>France</COUNTRY>
      </LOCATION>
      <ISSUEID>
        <DATE flux='R'>1998</DATE>
        <ISSUENO flux='R'>159</ISSUENO>
      </ISSUEID>
      <ISSN ISDS='Y' flux='R'>0245-9442</ISSN>
    </SERPUBFR>
  </SERFRONT>
  <FRONT type='PC'>
    <TITLEGRP>
      <TITLE TITYPE='1' LNGCODE='fre' ALPHABET='LATIN' flux='R'>Le
        ma&iuml;s transg&eacute;nique: Exemples
        d'int&eacute;r&ecirc;t agronomique</TITLE>
    </TITLEGRP>
    <AUTHGRP>
      <AUTHOR>
        <SURNAME flux='R'>GALLAIS</SURNAME>
        <FNAME flux='R'>A.</FNAME>
      </AUTHOR>
      <CORPAUTH>
        <ORGNOME flux='R'>INRA-UPS-INA-PG Station de
          G&eacute;n&eacute;tique v&eacute;g&eacute;tale</ORGNOME>
      </CORPAUTH>
    </AUTHGRP>
    <PUBFRONT>
      <PAGE>
        <FPAGE flux='R'>41</FPAGE>
        <LPAGE flux='R'>48</LPAGE>
      </PAGE>
      <EXTENT flux='R'>8</EXTENT>
    </PUBFRONT>
  </FRONT>
  <ABSTRACT lngcode='FRE' flux='P'>
    <P>Les travaux sur la transg&eacute;n&egrave;se du ma&iuml;s sont en
      plein d&eacute;veloppement aux Etats-Unis. La
      r&eacute;sistance aux insectes et la r&eacute;sistance aux
      herbicides sont des exemples de caract&egrave;res obtenus par
      g&eacute;n&eacute;tique. Des travaux sur la
      qualit&eacute; du grain sont en cours, ainsi que sur la
      modification du syst&egrave;me de reproduction. Le ma&iuml;s
      se pr&eacute;sente comme un cas particulier plut&ocirc;t
      favorable en mati&egrave;re de transg&eacute;n&egrave;se.</P>
  </ABSTRACT>
  <ACDOC>
    <CARACDOC>
      <CONDAC>
        <AVAIL flux='P'>MAG1</AVAIL>
        <ADEDOC flux='R'>24368</ADEDOC>
      </CONDAC>
    </CARACDOC>
  </ACDOC>

```

</ARTICLE>

Note that the user cannot access directly BIBL database. Indeed, it is not a classical bibliographical database because its constitution does not guarantees its exhaustiveness. BIBL will receive only secondary queries coming from DIS or ORD databases (this particularity is materialized by the dotted arrows between these databases in the schema).

3.3. The library management system

The data we obtain from LM permit essentially to complete the characterisation of the bibliographic notices coming from [Article@INIST](#) like, by example, the main scientific domains related to the each journal concerned by a document order or a bibliographic notice displayed. These data are extracted under request and go to enrich the records of a [Miri@d](#) proper bibliographic database, named BIBL in the schema.

3.4. [Article@INIST](#) log-files

The data obtained from the log-files of [Article@INIST](#) provide information about user behaviour facing a thought process of information retrieval in a scientific, multidisciplinary database. Pragmatically, we have data about both users queries and bibliographic references displayed identification. The extraction is done daily and [Miri@d](#) includes a number of modules to extract the information from the log-files of [Article@INIST](#) server and to organize it into sets of records and indexes. In these log-files, every request from the user is recorded with different kinds of information depending on the type of action performed.

For example, a search for an article looks as follows (some information about user identification and query contents were deleted for reason of confidentiality):

```
TRACE |PID:00255|20/07/1999|00:51:58|MESSAGE:a_search started
TRACE |PID:00255|20/07/1999|00:51:58|MESSAGE:Ouverture des collections de
: /artic/index/art_archive.clm
TRACE |PID:00255|20/07/1999|00:51:58|MESSAGE:Ouverture de la base :
/artic/index/art_1992
TRACE |PID:00255|20/07/1999|00:51:58|MESSAGE:Ouverture de la base :
/artic/index/art_1993
TRACE |PID:00255|20/07/1999|00:51:59|MESSAGE:Ouverture de la base :
/artic/index/art_1994
TRACE |PID:00255|20/07/1999|00:51:59|MESSAGE:Ouverture de la base :
/artic/index/art_1995
TRACE |PID:00255|20/07/1999|00:51:59|MESSAGE:Ouverture de la base :
/artic/index/art_1996
TRACE |PID:00255|20/07/1999|00:51:59|MESSAGE:Ouverture de la base :
/artic/index/art_1997
TRACE |PID:00255|20/07/1999|00:51:59|MESSAGE:Ouverture de la base :
/artic/index/art_1998
TRACE |PID:00255|20/07/1999|00:51:59|MESSAGE:Ouverture de la base :
/artic/index/art_1999
TRACE |PID:00255|20/07/1999|00:51:59|MESSAGE:Search for match
TRACE |PID:00255|20/07/1999|00:52:18|MESSAGE:(18706 ms) Elapsed Retrieval
TRACE |PID:00255|20/07/1999|00:52:18|MESSAGE:Display informations
TRACE |PID:00255|20/07/1999|00:52:18|MESSAGE:1
TRACE |PID:00255|20/07/1999|00:52:18|MESSAGE:10
TRACE |PID:00255|20/07/1999|00:52:18|MESSAGE:qstat: [128.**.**.**/
**.**.**.edu.au/] a_search/100/115/3306350/(18706 ms) Elapsed
Retrieval/** and ****<IN>AUTHOR <AND> (français)<IN>LANGUAGE
TRACE |PID:00255|20/07/1999|00:52:18|MESSAGE:Read the score and fields
TRACE |PID:00255|20/07/1999|00:52:19|MESSAGE:Display the list
TRACE |PID:00255|20/07/1999|00:52:23|MESSAGE:a_search terminated
```

This set of data, obtained for each query, permits to constitute one record in the [Miri@d](#) database named QUE in the schema.

Successive queries and visualizations from the same user (actually, from the same IP number) are regrouped in chronological order and then they are split in “sessions” following a simple heuristics. This allows to determine how many records were displayed after a given successful query.

In [Article@INIST](#) log-files we also have information about visualization of bibliographical records. The visualization of a bibliographical record gives:

```
TRACE |PID:19531|20/07/1999|00:07:14|MESSAGE:views_doc started
TRACE |PID:19531|20/07/1999|00:07:14|MESSAGE:Client de type :
[Inconnu/Inconnu/Inconnu]
TRACE |PID:19531|20/07/1999|00:07:14|MESSAGE:Notice :
[Inconnu#/artic/data/00/15/66/04/00156604.mono]
TRACE |PID:19531|20/07/1999|00:07:14|MESSAGE:fichier [/var/tmp/aaaa19531]
TRACE |PID:19531|20/07/1999|00:07:14|MESSAGE:views_doc terminated
```

Obtained for each bibliographical record displayed, this set of data forms one record in the [Miri@d](#) database named DIS in the schema.

4. [Miri@d](#) databases

The different databases are made of SGML (Standard Generalized Markup Language – ISO 8879:1986) records stored and retrieved in a special file system using programmes from the Ilib library.

4.1. QUERies database

This database is constituted by records describing the [Article@INIST](#) users queries. QUE database permits only primary queries concerning exclusively its own fields. Secondary queries to other [Miri@d](#) databases are forbidden.

The table 4.1 gives the database fields and their description. The nine fields written in *italic* characters produce indexes.

Field	Description
<i>Number</i>	sequential key identifying the query
<i>Domain</i>	user associated TLD (top level domain): country code (ISO 3166) or domain
<i>Date</i>	day and hour of the query
<i>Query mode</i>	query mode employed
<i>Year</i>	publication years asked
<i>Session</i>	query chronological position
<i>Query</i>	search fields activated by the query
<i>Results</i>	number of records obtained
<i>Display</i>	number of records obtained AND displayed
Total	total number of records inspected by the query

Occurrence	number of times same query was activated
------------	--

Table 4.1

4.2. DISplay database

The records of this database contain information about, in the one hand, bibliographical references identification and, on the other hand, users themselves. DIS database permits primary queries but also secondary queries to BIBL database.

The table 4.2 gives the database fields and their description. The four fields written in *italic* characters are indexed.

Field	Description
<i>Number</i>	sequential key identifying the record displayed
<i>Date</i>	day and hour of display
<i>Type</i>	type of the document corresponding to the record displayed
<i>Domain</i>	user associated TLD (top level domain): country code (ISO 3166) or domain
Identification	numerical key identifying the record displayed

Table 4.2

4.3. ORDers database

This database is composed by records containing, in the one hand, bibliographical references identification and, on the other hand, users themselves. ORD database permits primary queries but also secondary queries to BIBL database.

Database fields and their description are presented in the table 4.3. The four fields written in *italic* characters are indexed.

Field	Description
<i>Number</i>	sequential key identifying the record
<i>Date</i>	order year
<i>Location</i>	customer country code (ISO 3166)
<i>Activity</i>	customer activity code (INIST internal coding)
Identification	numerical key identifying the document ordered
Customer	customer code (not displayed control data)
Orders	total number of customer orders delivered in year
Turnover	turnover corresponding to customer orders delivered

Table 4.3

4.4. BIBLIographical database

This database contains information about the content of bibliographical references displayed or corresponding to an ordered document. User does not directly access to this database. BIBL database receives only secondary queries coming from DIS or ORD databases.

The table 4.4 gives the database fields and their description. The nine fields written with *italic* characters are indexed.

Field	Description
<i>ISSN</i>	key, done by the ISSN International Centre, identifying the journal or monograph related to the bibliographic reference
<i>Title</i>	journal title done by ISSN International Centre
<i>Language</i>	text language
<i>Country</i>	publishing country
<i>Paper</i>	paper original title
<i>Translation</i>	paper title translated into English (if original title in no-Latin characters)
<i>Author</i>	paper authors
<i>Year</i>	paper publishing year
<i>Scientific</i>	key giving one or more scientific domain related to the paper
Reference	key identifying the journal issue and the publication related to the bibliographic reference (this key is a concatenation of two numerical keys done by INIST)
Periodical	key identifying the journal or monograph related to the bibliographic reference (this key is done by INIST)
Short	abbreviation of journal title done by ISSN International Centre
Publisher	publisher name
Volume	volume, belonging to a publishing year, where the paper appears
Number	issue, belonging to a volume, where the paper appears
Pages	first and last pages numbers of the paper
Total	paper number of pages

Table 4.4

4.5. STATistics database

In the schema we have a fifth [Miri@d](#) database: STAT database. These database proposes a simplified, direct access to some a priori calculated statistics indicators based on the data stored in the other [Miri@d](#) databases: QUE, DIS, ORD and BIBL. The user must express his preferences choosing a periodicity, then a period and finally an indicator. The indicators are calculated off-line with a periodicity of one day, one week, one month or one year according to the available data characteristics by shell scripts and/or Perl programmes using programmes from Ilib library.

Table 4.5 shows a description of proposed indicators:

Periodicity	Indicator
D/W/M	distribution of successful queries by query mode
D/W/M	distribution of unsuccessful queries by query mode
D/W/M	distribution of displayed records after one query by query mode
W/M/Y	distribution of queries by date
D/W/M	total number of displayed records
D/W/M	distribution of displayed records by bibliographic type
W/M/Y	distribution of displayed records by date
W/M/Y	first $\{N_w, N_m, N_y\}$ records most displayed
W/M/Y	distribution of users by geographical origin or TLD domain
W/M/Y	first $\{N_w, N_m, N_y\}$ journals most displayed
W/M/Y	distribution of users by geographical origin or TLD domain
W/M/Y	distribution of users by geographical origin or TLD domain, by journal
W/M/Y	distribution of most displayed journals by publishing country
Y	total number of ordered documents
Y	distribution of customers ordering by country
Y	first N_y documents most ordered
Y	distribution of customers by geographical origin
Y	distribution of customers by activity
Y	distribution of ordered documents by publishing year
Y	first N_y journals most ordered
Y	distribution of customers by geographical origin, by journal
Y	distribution of customers by activity, by journal
Y	distribution of ordered journals by publishing country

Table 4.5

5. [Miri@d](#) statistical module

MIRI@D includes a number of modules in order to apply standard descriptive statistics to different retrieved records and Web user searches. MIRI@D produces these descriptive statistics on the retrieved records contained in the databases and on the Web user searches.

The analyst uses a browser in order to assemble a corpus of records from Web user searches, but also to define a descriptive statistical analysis. Each type of descriptive statistical analysis is performed using scripts and/or Perl programmes using programmes from Ilib library.

The results are available to analysts, who can use this descriptive statistical information as a source for their indicator design tasks, and as input for data analysis. Results can also be useful to managers, in order to improve management and decision-making.