ECP 2006 DILI 510049

ENRICH

Report on search behavior of Manuscriptorium users

Deliverable number	*D 4.2*
Dissemination level	*Public*
Delivery date	*6.5.2009*
Status	*Final*
Author(s)	*Michal Zyka, Martin Majer, Tomas Psohlavec, AIP*

e**Content*****plus***

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

Document Version Control

Version	Date	Change Made (and if appropriate reason for change)	Initials of Commentator(s) or Author(s)
1.0	22/04/2009	Initial Version	TP, AIP
1.0.1	29/04/2009	Reviewed version	MICF (R. Caldelli and R. Becarelli)
1.0.2	06/05/2009	Updated accordingly to review suggestions, graphs added	TP, AIP

Document Review

Reviewer	Institution	Date and result of the review
Roberto Caldelli	MICF	29.4.2009. Improvement of chapter 2 suggested.
Jakub Heller	CCP	Document approved for submission to EC

Approved By (signature)	**Date**

Accepted by at European Commission (signature)	**Date**

1. Executive Summary

Deliverable D4.2 reports the user search behavior within the Manuscriptorium routine service. The results are interpreted on a basis of a log file created by the retrieval engine and apply to month 10 – month 17 of duration of the ENRICH project. The results are interpreted from various points of view and a final conclusion is attached indicating the focus of further development tasks in WP4.

CONTENT

2. Searching in Manuscriptorium & log file

The Manuscriptorium search engine uses a set of indexed fields to perform search tasks. The indexes are filled with certain information from the XML records on the basis of an index-dedicated XSL transformation. This solution enables us to create “virtual” search fields especially for search purposes and these fields are filled with information originated from various locations in the original XML.

To illustrate the possibilities of preparation of search fields we include here the following example:

Example field name	Field content description	Field content origin
Author	All names of persons with primary author’s responsibility related to the source document	Contents of all `author` elements located within any `msItem` elements located within `msDesc` using XPath: each `/descendant::msDesc/descendant::msItem/author`

Person with intellectual responsibility	Content of author field + all other intellectual responsibilities	Contents of all `author` elements located within any `msItem` elements located within `msDesc` and all `name` elements with personal names located directly within `respStmt` element anywhere within `msContents`
		
		using XPath:
		each
		`/descendant::msDesc/descendant::msItem/author`
		and each
		`/descendant::msDesc/msContents/descendant::respStmt/name [@type="person"]`

Name	All personal names without any limitations	All `name` elements with personal names and all `author` elements
		
		using XPath:
		each
		`/descendant::author`
		and each
		`/descendant::name[@type="person"]`

Date of origin	All origin related dates	All `origDate` elements in `msDesc/head` element and `date` elements in `msDesc/history/origin` elements and `origDate` elements in `msDesc/history/origin` elements

The artificial examples above illustrate the advantages of such approach:

1) the end-user does not need to specify exactly the search location(s) in the XML record (this would require detailed knowledge of the format structure which cannot be expected); he or she just indicates the type of information which should be searched: e.g. in case of Date of origin it is not necessary to know where all related information can be found.
2) The indexes for the searching are prepared in advance and the level of load at the server side is significantly decreased

The user can use three different forms (these are described later) to search such pre-prepared fields and it is possible to formulate very simple but also very sophisticated queries.

The particular behavior and the way of real usage of all the available features are monitored in a dedicated log file. This monitoring feature was implemented in the initial stage of the WP4.

2.1. Recorded information

The log file records every search query applied within the retrieval interface of the Manuscriptorium service. In other words this means:

- **field(s)** to be searched and the **query words/terms** to search form
- type of **search form** used to create a query
- **search options** applied
- **date** and **time** of the query
- **language** of interface

The log file is produced as an XML and is available for further processing.

The log file relates to searching of the catalogue of Manuscriptorium which contains (during the log recording period) approx. 180 000 records about historical documents. The recorded numbers are not related to accessing of digital documents within the Manuscriptorium digital library (approx. 4 000 documents during the recording period). The resources in the digital library can be accessed using a direct access (through portals and other ways of access) bypassing the research system of Manuscriptorium. Therefore the numbers recorded are used to interpret solely the end-users search behavior, they cannot be used to measure the degree of accessing the resources (records, documents) within Manuscriptorium platform.

2.2. *The purpose of the log file*

Answers to the following important questions were intended to be found using the log file:

- how effectively is the set of available fields used?
  - which fields are used the most and how these most important fields content could be improved?
  - which fields are not used - should we improve the content or remove them as unnecessary?
- how effectively are the different forms (quick query form, simple form, advanced form) used?
- which tasks should be focused during further development?

3. Interpretation of results

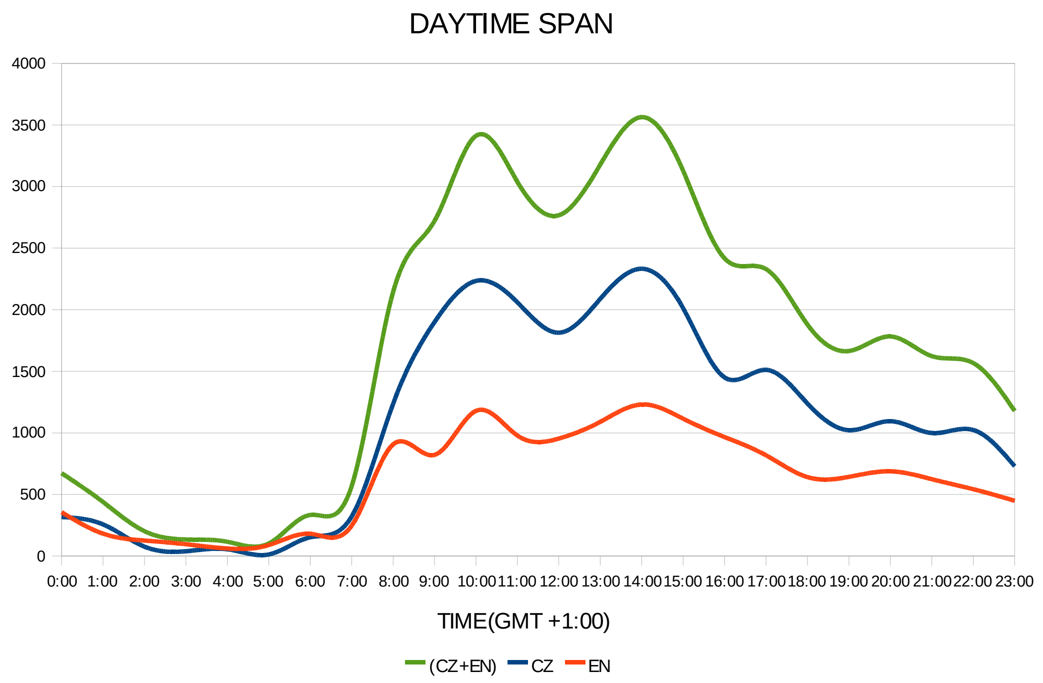
3.1.1 General information

Total number of queries

The users laid 40 962 queries during the monitoring period.

3.1.2 Daytime usage

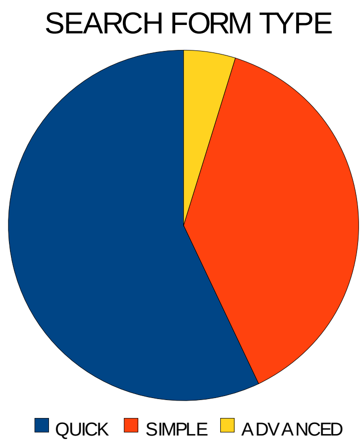
The following picture illustrates the average frequency of queries during a day.



The highest Manuscriptorium retrieval system usage rate is definitely between 8:00-11:00 and also 13:00-17:00 with mild decrease in the noon (lunchtime). The peaks at 10:00 and 14:00 indicate that the retrieval system is used during work time and school time.
The lower usage rate in the morning and during night reflects the fact that the Manuscriptorium is accessed mainly by the Europe based users.

3.2 Query forms and query complexity

The following picture shows the usage of the three available types of search forms.

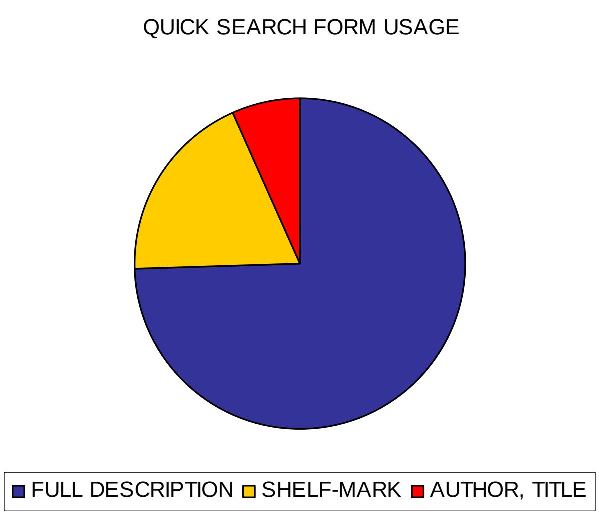


The results are as follows:

- Quick Search form: 57 % of all queries
- Simple Search form: 38 % of all queries
- Advanced Search form: 5 % of all queries

3.2.1 Quick search form

The Quick Search form is located directly in the homepage of Manuscriptorium website and its main purpose is to provide the quickest access to desired records. As seen above it is the most often used way of searching in Manuscriptorium. In relation to this search form we can see the following numbers in the log file:



- 67 % of the queries searched through full descriptions (without constraints to one of the available fields)
- 17 % of the searches were made directly via a known shelf-mark (using dedicated fields)
- the possibility to search the other two fields available in this form – the author and title field – used only a few of searches; the percentage ratio is approx. 6 % for each field

Conclusion:

The information above indicates that the form is used as originally intended. There is no need changing the available fields nor add any specific search settings as the users mostly do not use even the basic available fields. When using a Quick Search form they require either

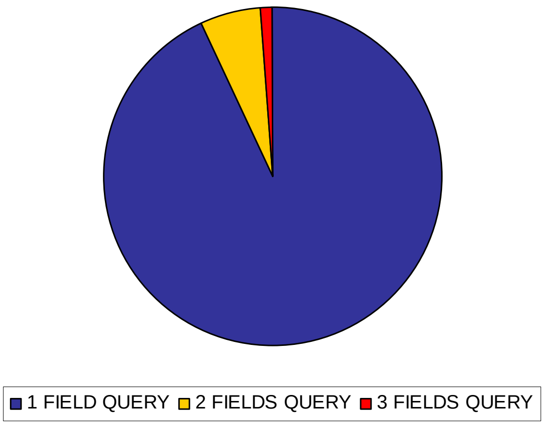
- a general wide results matching a simple expression
- a particular document matching a known shelf-mark

3.1.2 Simple and Advanced Search form

These forms are similar and both are primarily designed to enable queries combined above multiple fields. Each form allows combining of queries above up to three fields using boolean operators. The differences between those two forms are:

- the Simple Search form has two predefined fields (Settlement + Repository fields), the set of the fields to be searched in the Advanced Search form is freely selectable
- there is a fixed operator AND used to indicate the relation between the three fields in the Simple Search form, the operators in Advanced Search form are freely selectable
- the various search options are either predefined or disabled in the Simple Search form (so the results of searches are wider) while in the Advanced Search form the user has the possibility to fully adjust the settings and control the search behavior of the retrieval system – therefore more exact results can be expected

The following numbers illustrate how many queries really use the possibility to combine queries above multiple fields:



- 1 field query: 93 % of all Manuscriptorium queries
- 2 field query: 6 % of all Manuscriptorium queries
- 3 field query: 1 % of all Manuscriptorium queries

The following numbers are also interesting:

- only 18 % of all queries created within Simple or Advanced Search forms really uses the possibility to combine queries above 2 or 3 fields.
- only 3 % of all queries created within Simple or Advanced Search forms really uses the possibility to combine queries above 3 fields.

There is also a possibility to use checkboxes to constrain the search in certain ways e.g. limit the search to certain document types, to documents with digital copy, to documents with fulltexts etc. The research system invisibly extends the query by additional field constraint. Such easy-to-use constraints were attached to:

- 13 % of all Manuscriptorium queries
- 9 % of queries applied in the Quick Search form
- 21 % of queries applied in the Simple Search form
- 1 % of queries applied in the Advanced Search form

These easy-to-use features are not included in the Advanced Search form (but it is possible to use dedicated field instead) and are fully available in the Simple Search form only (Quick search includes only one such checkbox). It is obvious that users of Manuscriptorium prefer such easily accessible predefined queries (query parts) rather than creating the queries accurate by combining various fields.

Conclusion:

We suppose that the ratio between application of the three types of forms indicates that users are interested in simple forms, possibly willing to use forms with predefined queries. There is no need or no will to use a more “technical” approach to searches as it is represented nowadays by the Advanced Search form in Manuscriptorium.

This assumption is in a good conformance with the ratio between queries above one, two and three search fields.

In contradiction the frequency of use of predefined query part with multiple search fields indicates a possible way of improvement of the end-user interface and way to make the query results more accurate.

3.3. Fields usage

During the monitoring of search behavior we focused on the most frequently used fields. There are following fields currently available in the Manuscriptorium routine service (sorted according to the frequency of use within search queries):

- Full record — 49 % of all queries
- Shelf-mark: — 15 % of all queries
- Is digital document available?: — 11 % of all queries
- Settlement: — 6 % of all queries
- Author: — 6 % of all queries

- Title: 5,5 % of all queries
- Repository: 5 % of all queries
- (Is fulltext available?), Date of Origin, Date, Incipit, Country, Name, Provenance, Printer: approx. 3 % of all queries
- Bibliography, Alternative Name, Place of Origin, Origin, Rubric, Music Notation, Scribe, Responsibility - Name, Additions, Explicit, Colophon, Responsibility: approx. 0,5 % of all queries

Conclusion:

The seven most frequently used items in the list match exactly with the fields available within the Quick Search and Simple Search forms. We can assume that the users use exactly the fields which are the most visible.

4. Conclusion and indications for further improvement of the retrieval interface

Considering the partial conclusions we can say that:

- end-users create very simple queries
- end-users often access particular documents they already know
- the end-users prefer to use predefined query parts when they want to receive more accurate results
- the potential of the retrieval engine itself is not exploited by the end-users

This can be due to:

- the interface for creating accurate queries is too difficult to use or hidden from the end-users
- the wide results are still satisfying enough for the end-users
- the set and contents of available fields does not match the needs of target end-users so they use the most basic fields only

Therefore we should focus on the following issues during further development:

- make the retrieval interface easier to use, with focus to
    - improvement of Quick Search and Simple Search forms (abandonment of the current Advanced form approach)
    - inclusion of predefined constraints in some easily understandable form
    - substitution of search above more fields with better utilized feature of subsequent queries (e.g. making the results as accurate as necessary in sequence of simple queries)
- verify or re-design the available set of fields and the way these are filled with contents
- there is no need to improve the retrieval engine core, all the changes should be done in the upper levels of end-user interfaces.

These results should be taken into account when resolving T4.4 and also T4.5 related to the Manuscriptorium search engine.