ENRICH PROJECT

ECP 2006 DILI 510049

ENRICH

Report on definition of typical model searches and their implementation

Deliverable number	*D 4.4*
Dissemination level	*Public*
Delivery date	*02/12/2009*
Status	*Final*
Author(s)	*Paolo Mazzanti, Roberto Caldelli, MICF and all WP4 partners..*

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	09/11/2009	Initial version.	PM, MICF RC, MICF
2.0	16/11/2009	Second version	PM, MICF TP, AIP RK, UZK FW, UZK
3.0	19/11/2009	Pre-final version	RC, MICF
3.1	02/12/2009	Final version	RC, MICF

Document Review

Reviewer	Institution	Date and result of the review
Tomas Psohlavec	AIP	10/11/2009: The structure of the document approved, updates related to the latest improvements of the searching engine included.
Jakub Heller	CCP	02/12/2009: Formal check and formatting control of the document before submission.

Approved By (signature)	Date

Accepted by at European Commission (signature)	Date

1 Executive Summary

Deliverable 4.4 reports the activity that has been carried out within Task 4.5 of WP4 (User personalization) and describes the functions and solutions of the typical model searches and their implementation in the new Digital Library implemented in Manuscriptorium. In particular, D.4.4 concerns the implementation of deep searching possibilities above all metadata and textual data, and the access to the digital contents through intelligent operators, to allow the user to easily achieve the requested results. Furthermore, a prototype which implements basic semantic search facilities has been designed and realized.

CONTENT

2 Introduction

The database systems generally allow for access to stored data for the widest possible spectrum of users, with effort of the simplest search method to achieve the highest precision of the requested information. Within ENRICH some of the subtasks focus this area, one of them more precisely focuses „the access to the digital content through intelligent operators, performing a semantic search of metadata“.

The following conception describes an alternative usage of the current Manuscriptorium search engine with a view to a semantic search. This document describes the solution achieved within Manuscriptorium considering the available tools in the Manuscriptorium platform, ENRICH timeframe and also other sources available within ENRICH project.

From the homepage of the new interface of Manuscriptorium http://beta.manuscriptorium.com/ (fig.1) the user has the access to the Digital Library where there are many possibilities to search the database. As it is written in the homepage, the user can choose the search method he/she prefers:

- Easy Search
- Advanced Search
- By Document Identification
- By Document origin

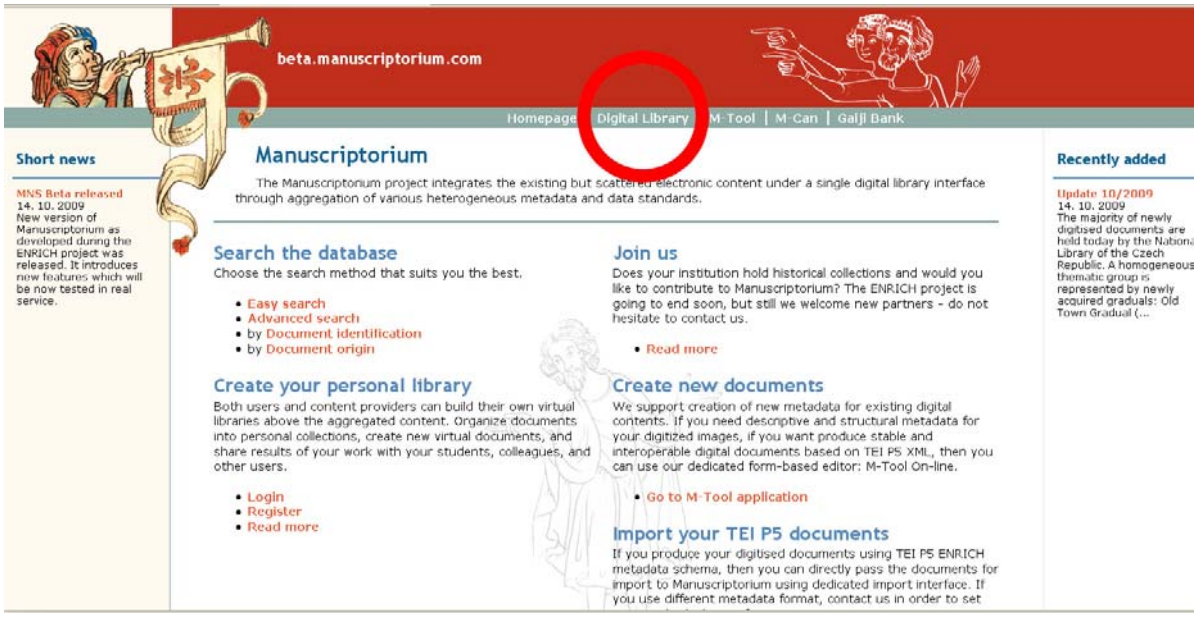

Fig. 1

In particular, in this report we focused on the new implementation of the “Easy search” and on the possibility to prepare more or less complicated queries in the form of predefined sentences, which can be combined in order to easily create precisely focused queries.

By clicking the "to Easy" link in the right sidebar of the Digital Library (Fig. 2), the user has the access to the Easy Serach form (Fig. 3).

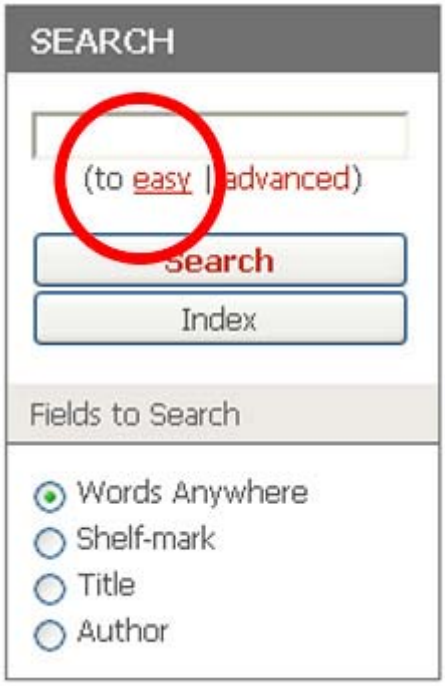

Fig. 2

3 Basic Search

The basic sentence creation uses predefined sentences which already contain some more or less complicated queries behind. The end-user does not have to care for all the boolean/intelligent/morphology operators nor particular search field structure behind the sentences. In fact, when using this search approach, such information is hidden to the end-user. Instead of that the natural semantic interpretation of the sentence should give an exact idea about the target of the intended search. With such a set of predefined sentences the „basic semantic search“ will be understandable and easy to use. Still it will be an efficient way how to retrieve the desired sets of documents.
The usability/accuracy of this search approach will depend on the usability/accuracy of the predefined query sentences. Within the pilot solution an easily understandable set of query sentences is available covering necessary scope of database queries and also correct interpretation of such queries in the search engine.

3.1 Combining sentences

Considering the possibility to combine the sentences into more sophisticated queries: in order to keep simplicity in the interface we have decided to enable to combine the sentences by placing more subsequent queries. E.g. only one sentence can be placed as a query at a time, the user receives search results as a response and then has a possibility to place another query extending or refining the last query results. The history of the search in the form of applied search sentences list is also maintained.

3.2 Finding of appropriate sentences

The sentences are originated as a result of discussion with end-users. Within ENRICH there are institutions managing historical sources and we can assume that they also have staff members experienced with the usual searching needs.

An open discussion regarding query sentences could be led regardless of what is the current set of search options provided by current Manuscriptorium search engine.

Finally the set of query sentences should be created regardless of what are the current searching possibilities of the Manuscriptorium interface and regardless the current Manuscriptorium end-users search behavior as the current behavior is implied by the current interface and current set of fields and for the new way of searching considered here we should not be limited by such inputs.

This approach can provide the best results. On the other hand it may result in having sentences/requirements impossible to apply within current Manuscriptorium. Should we later realize that implementation of certain search queries is not possible it will be a valuable feedback regarding the engine usability - it will indicate possible areas of additional search development.

4 Implementation

4.1 Set queries

First of all a set of different queries was defined. These queries are defined by an external XML file using specific rules for formulation of both query sentences and the final query behind - therefore the set of questions/queries is easily adjustable according to future user feedback.

The following screenshot (fig. 3) displays the actual set of available queries as it was generated for the “Easy search” interface (this is how the basic semantic search interface is now called within Manuscriptorium interface):

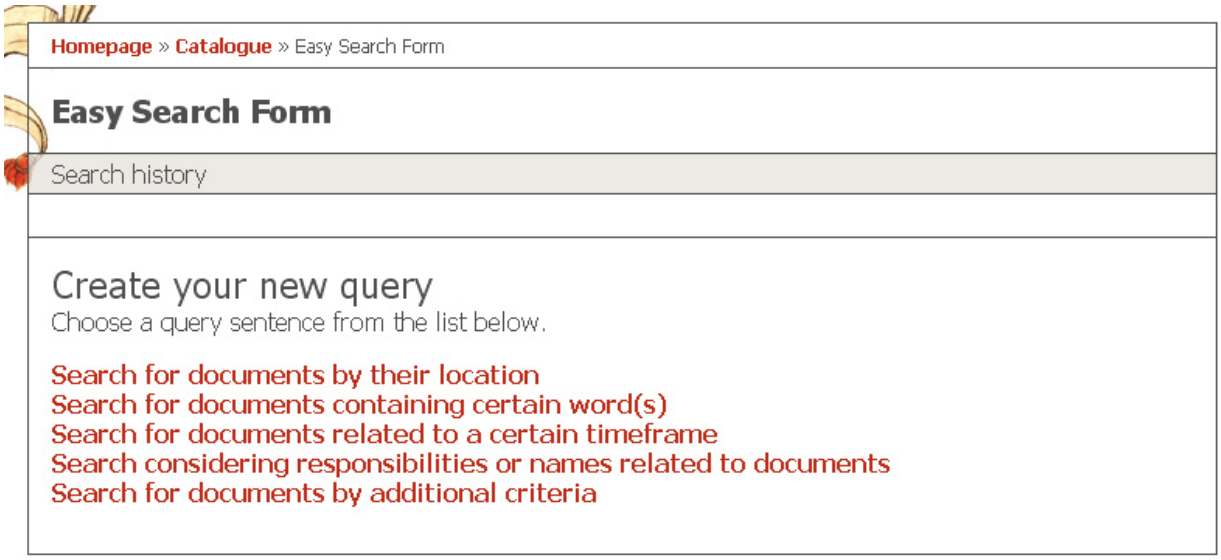
Homepage » Catalogue » Easy Search Form
Easy Search Form
Search history
Create your new query
Choose a query sentence from the list below.
Search for documents by their location
Search for documents containing certain word(s)
Search for documents related to a certain timeframe
Search considering responsibilities or names related to documents
Search for documents by additional criteria

Fig. 3

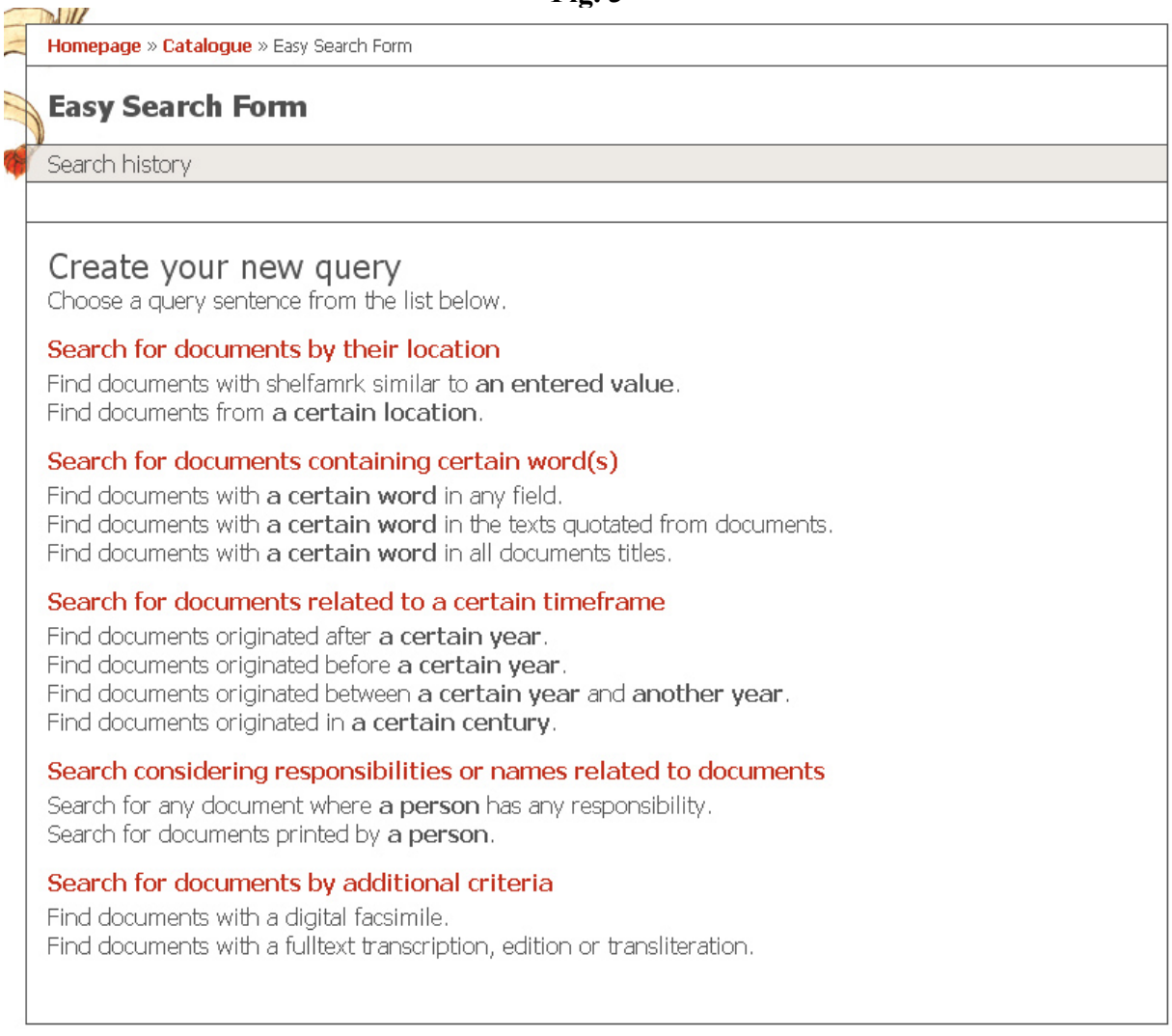
Homepage » Catalogue » Easy Search Form
Easy Search Form
Search history
Create your new query
Choose a query sentence from the list below.
Search for documents by their location
Find documents with shelfamrk similar to an entered value.
Find documents from a certain location.
Search for documents containing certain word(s)
Find documents with a certain word in any field.
Find documents with a certain word in the texts quotated from documents.
Find documents with a certain word in all documents titles.
Search for documents related to a certain timeframe
Find documents originated after a certain year.
Find documents originated before a certain year.
Find documents originated between a certain year and another year.
Find documents originated in a certain century.
Search considering responsibilities or names related to documents
Search for any document where a person has any responsibility.
Search for documents printed by a person.
Search for documents by additional criteria
Find documents with a digital facsimile.
Find documents with a fulltext transcription, edition or transliteration.

Fig. 4

As you can see (fig. 4) the sentences are organized into groups of thematically related queries. After clicking on the group header the group is extended or collapsed. The groups can be combined into a tree of sentences groups.

4.2 *Querying the database*

When reading the sentence the user gets the full meaning of the query and the expression that can be changed is emphasized.

For instance the user can choose a sentence "Find documents originated in **a certain century.**" in the „Search for documents related to a certain timeframe" group. It is obvious that there will be a possibility to select a certain century from an active list of choices.

This can be done by mouse click on appropriate sentence: after a sentence is selected the query is displayed in the upper part of the interface and the appropriate value can be set. Another example screenshot is available below (fig. 5):

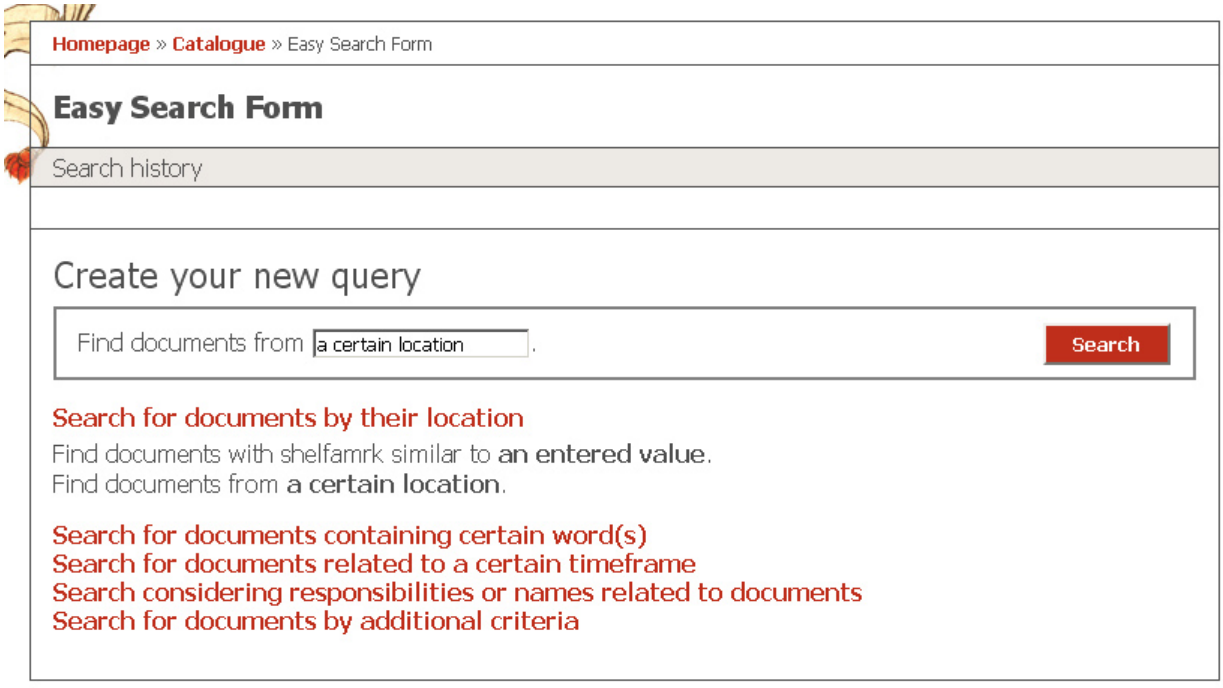

Fig. 5

The user can then enter, for instance, any "location" into the sentence. For instance a name of a town, country or even a library where the target documents are located can be entered. The system interprets the query as necessary.

After clicking to the Search button search results are obtained.

4.3 Search results and query sequences

While obtaining the search results the query sentence review is displayed in the "Search history" section. So the user can see what he/she asked for and also how many results were returned – see the screenshot below (fig. 6):

Fig. 6

When having some results of one or more preceding queries it is possible to continue the search by one of following ways (by clicking on the context links below the search history):

- Start a **new query**
- **Expand** results: the current results will be extended by results obtained after the next query (fig. 7)
- **Refine** results: the current results are refined by the next query (fig. 8)
- **Exclude** from results: records/documents corresponding to the next query will be removed from the current search results (fig. 9)

After a way of search continuation is selected the system reformulates the query sentences to naturally express what will happen with the search results – see the screenshots below:

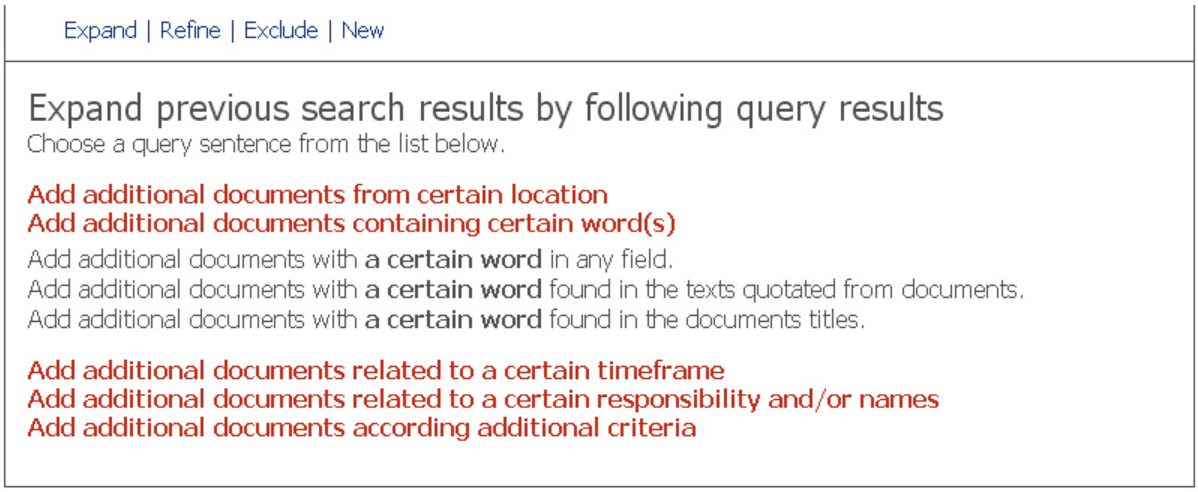

Fig. 7

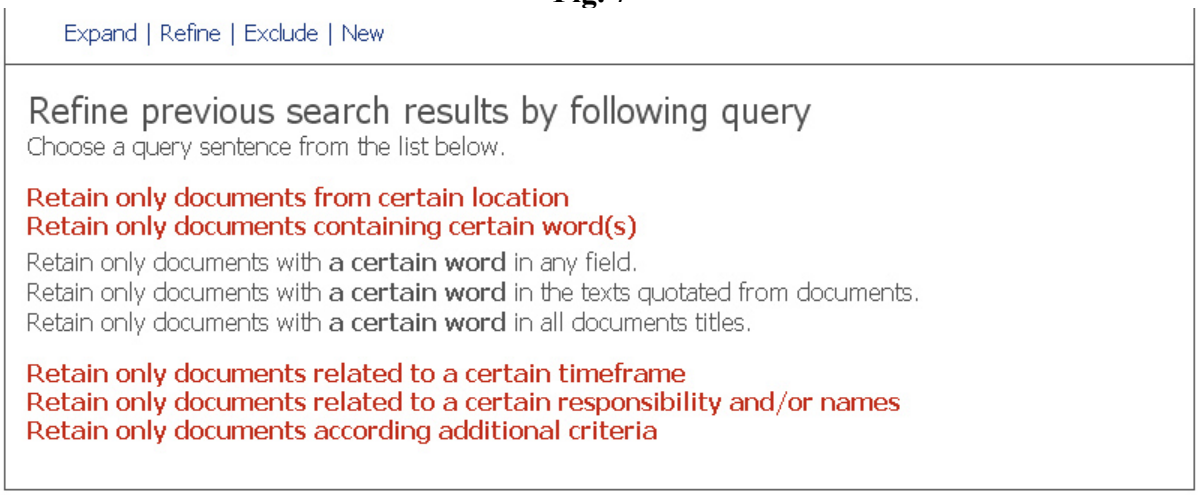

Fig. 8

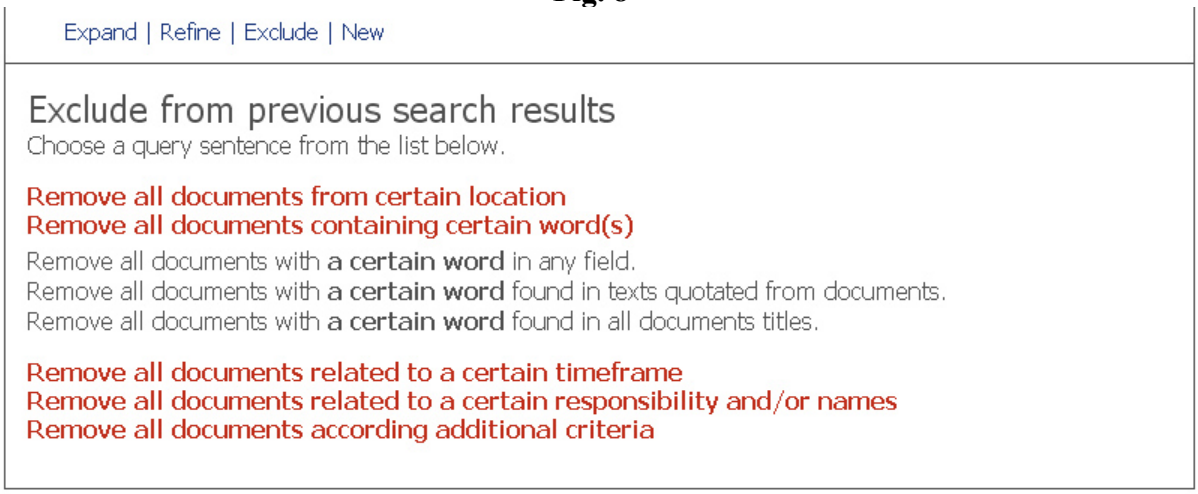

Fig. 9

The „Search history“ review again informs fully about what searches were performed. So it is possible to review the query sequence to the beginning of querying and also it is possible to return one step back to the previous query results (fig. 10).

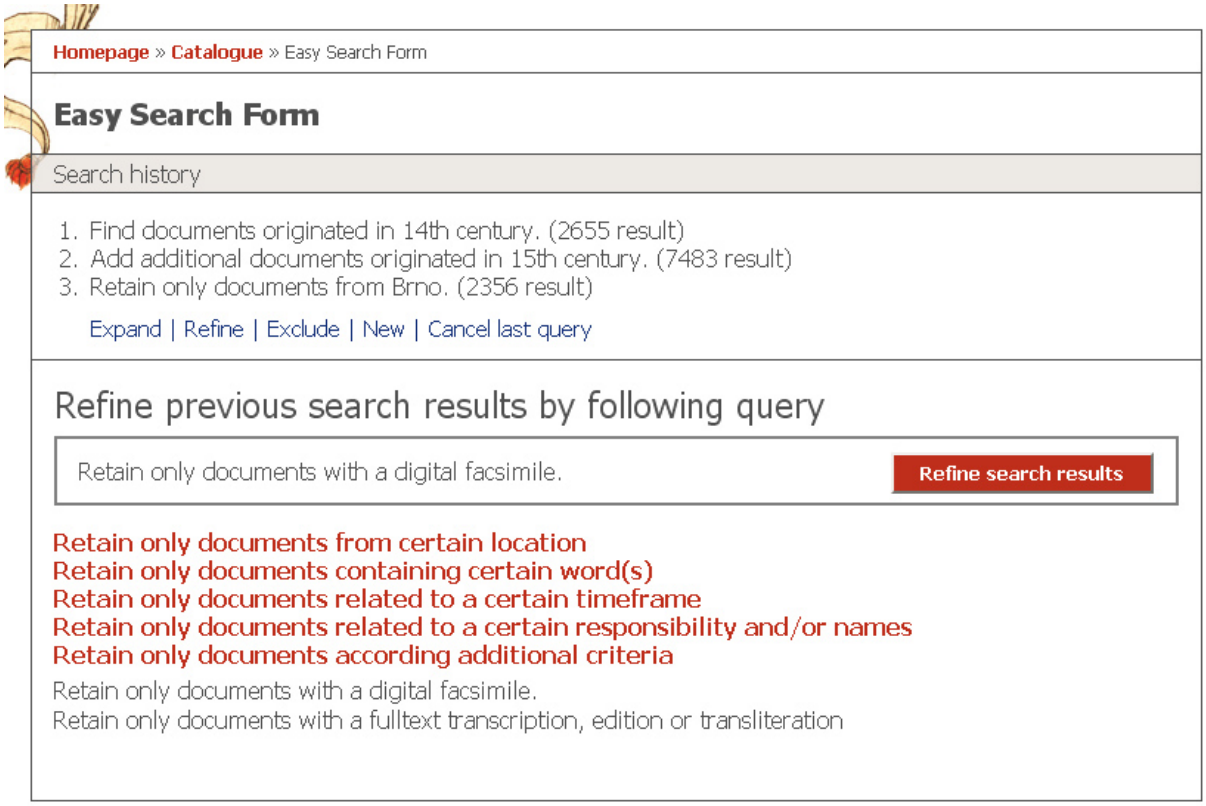

Fig. 10

4.4 A search use case

The following screenshots in this paragraph display a use case of "easy search" in the Digital Library. Inside the "Easy Search Form" the user may choose a query sentence (fig.11) from the list, in this case we decide to search documents by their location and specifically from a certain location where the documents are located. (fig. 12).

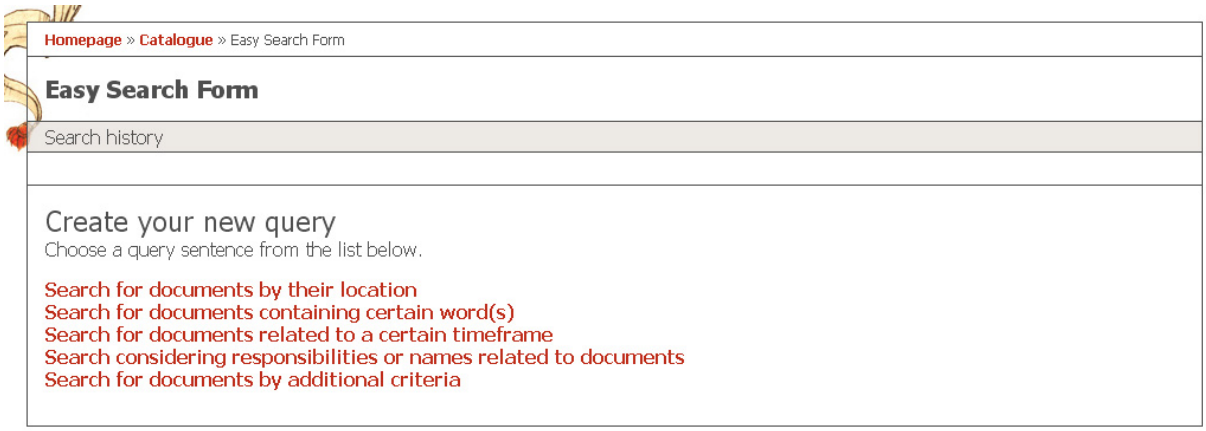

Fig. 11

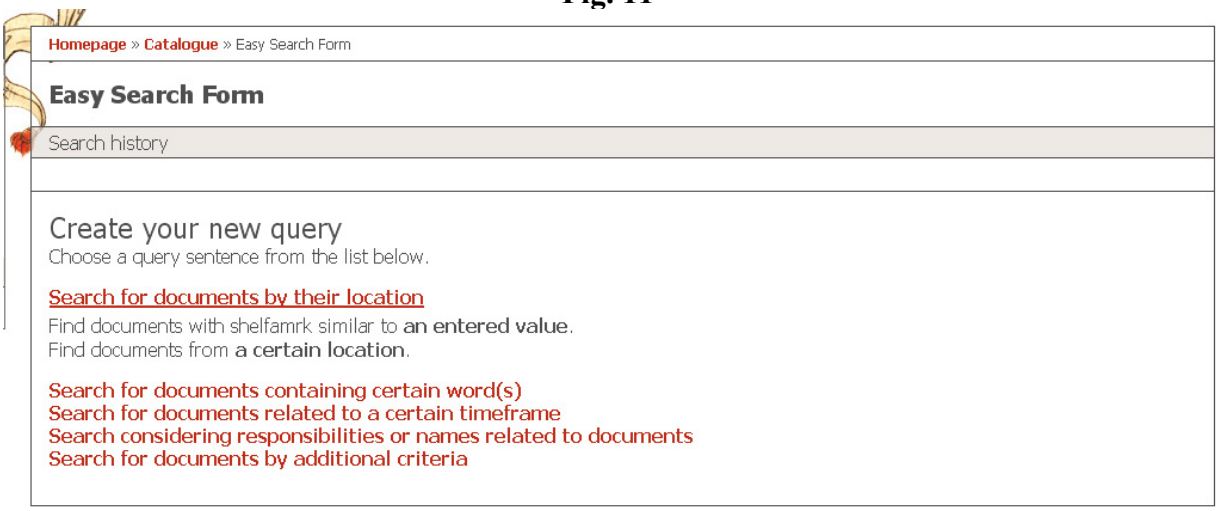

Fig. 12

In the search box we can digit the name location documents are searched from: in this case Praha (fig. 13). Clicking to the Search button search results are obtained.

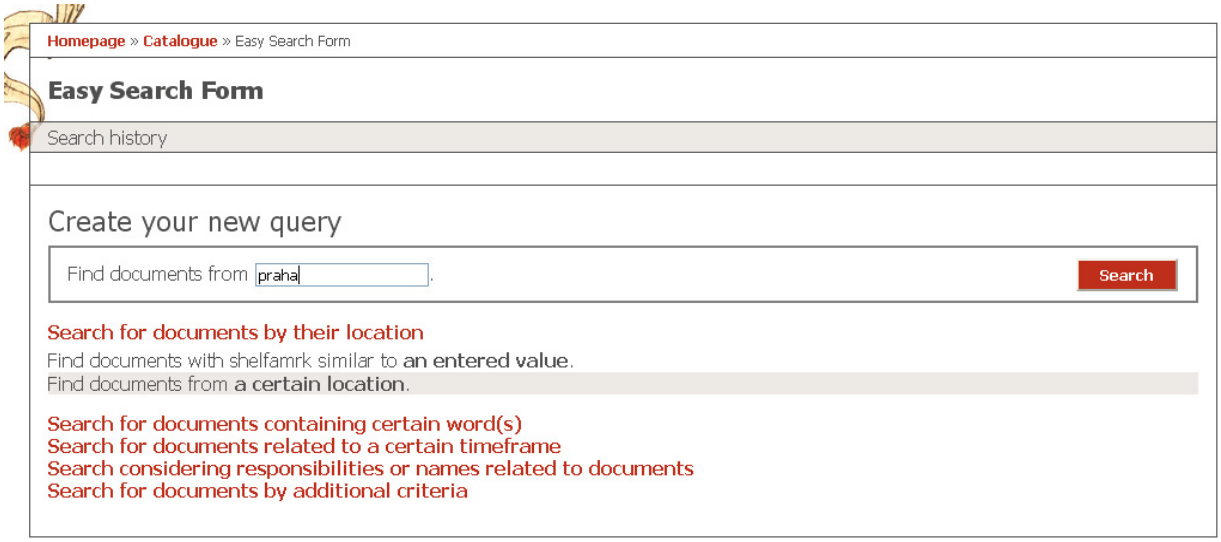

Fig. 13

The number of the records of this search appears on the top title Search results and the Search history shows at the point 1 that the documents from Praha are 2856 (fig 14).

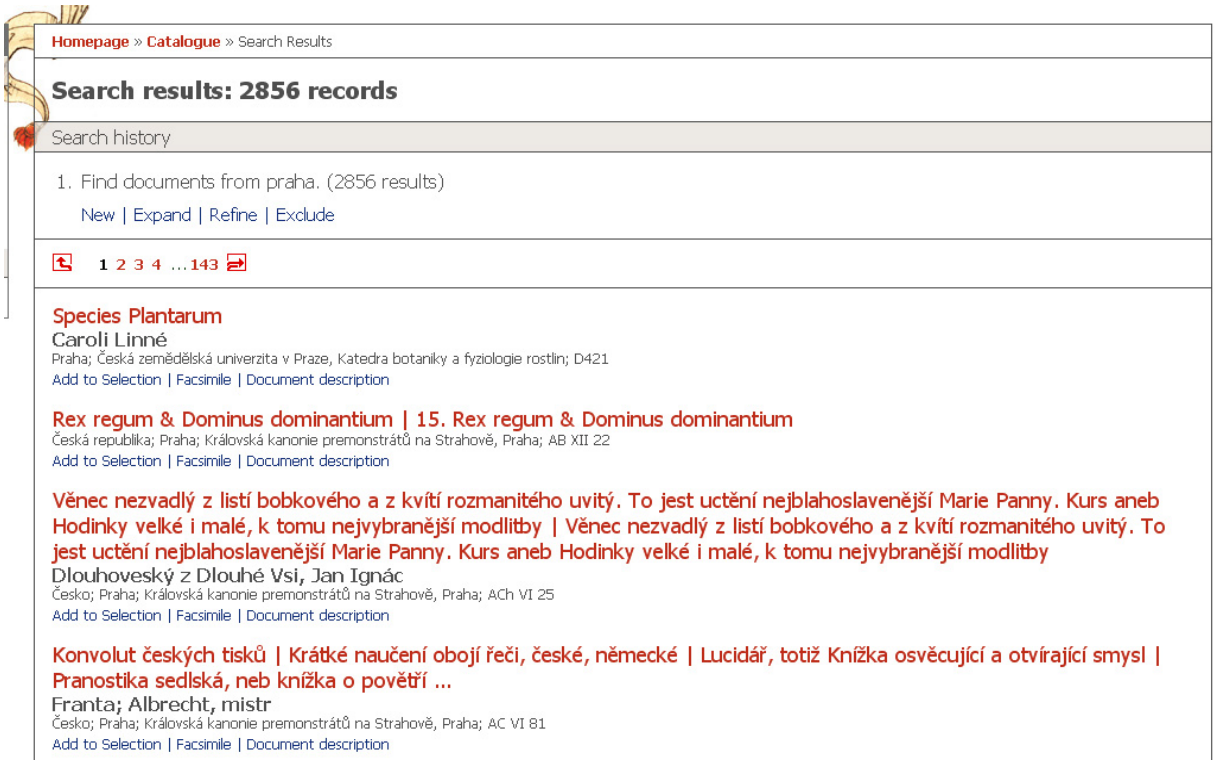

Fig.14

Now we continue to **expand** this search clicking on the context link “Expand” below the search history (fig. 15).

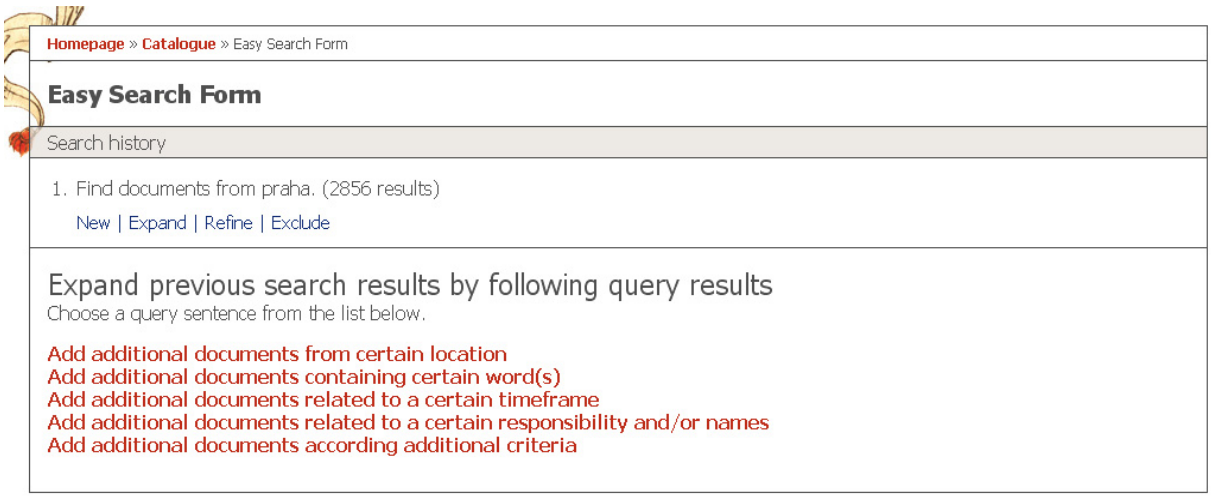

Fig 15

We are interested to add additional documents related to a certain timeframe, i.e. the documents originated in 17th century. Clicking to the Expand search button search results are obtained (fig. 16).

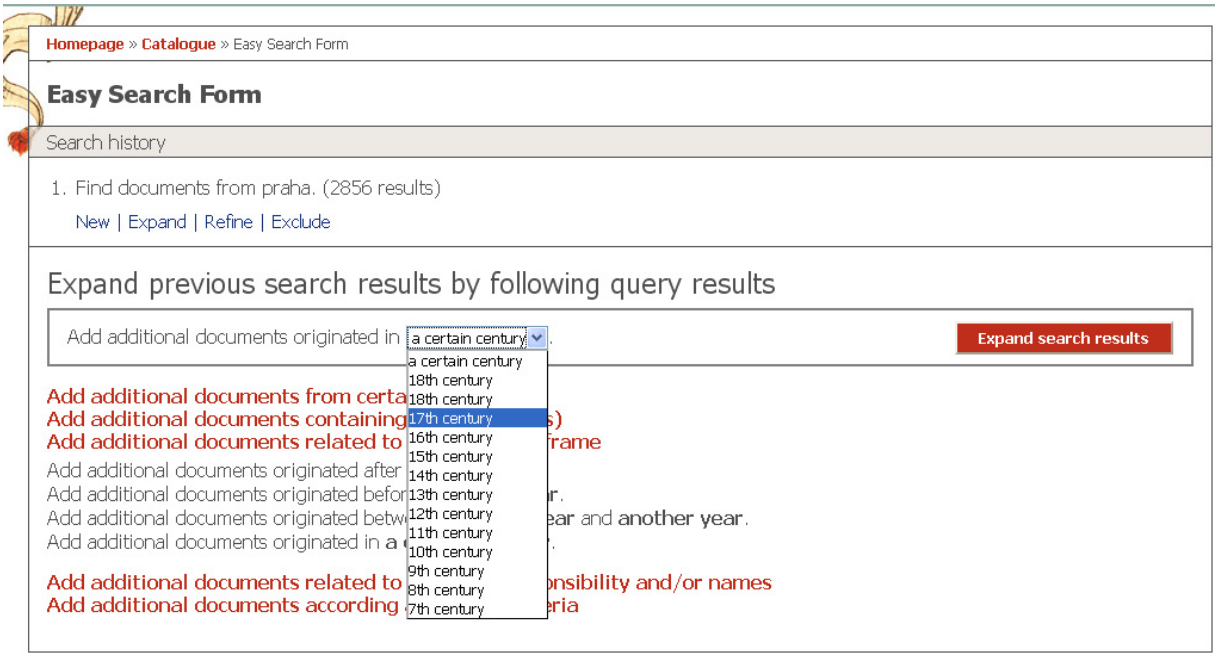

Fig. 16

The number of the records of this search appears on the top title Search results and the Search history shows at the point 1 the result of the first previous search and at the point 2 the result of the expanded search shows that the documents originated in 17th century are 5780 (fig. 17).

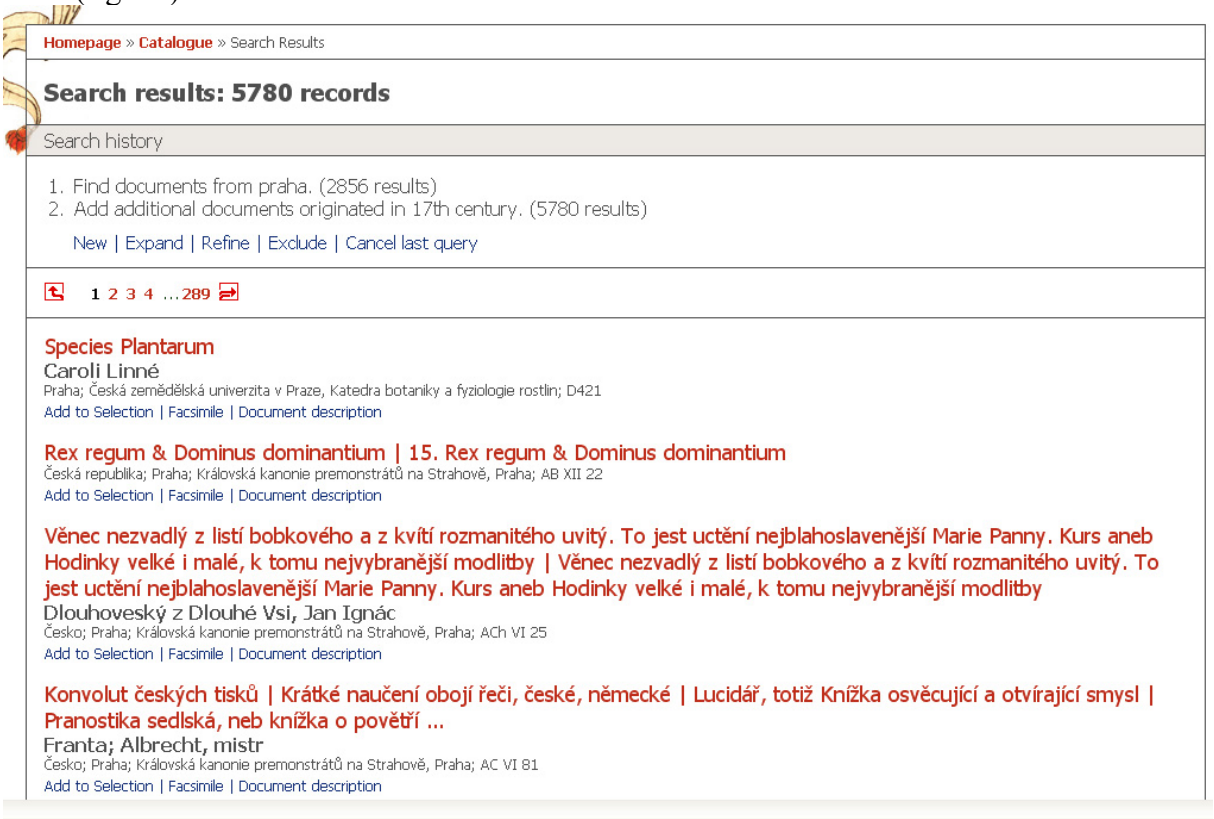

Fig 17

At this point, we continue to **refine** the search clicking on the context link "Refine" below the search history (fig. 18). We are interested to retain only documents with the *rex* word in any field.

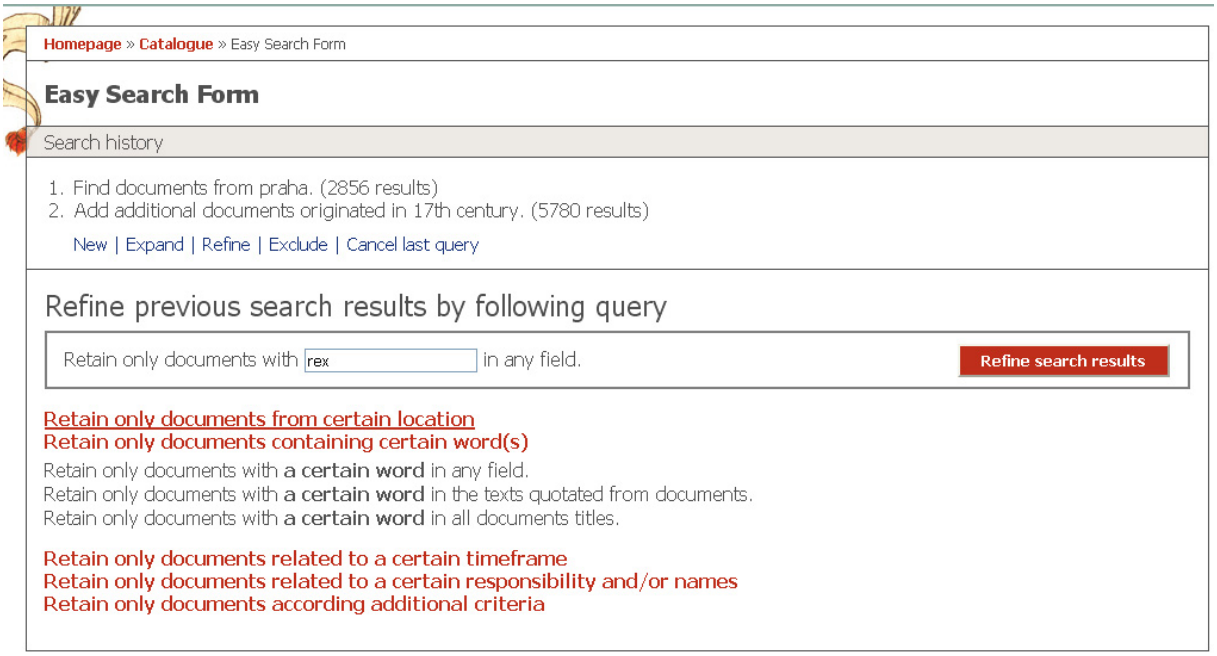

Fig. 18

The number of the records of this search appears on the top title Search results and the Search history shows at the point 1 and 2 the results of the previous searches and at the point 3 the result of the refined search shows that the documents with the *rex* word in any field are 149. (fig. 19).

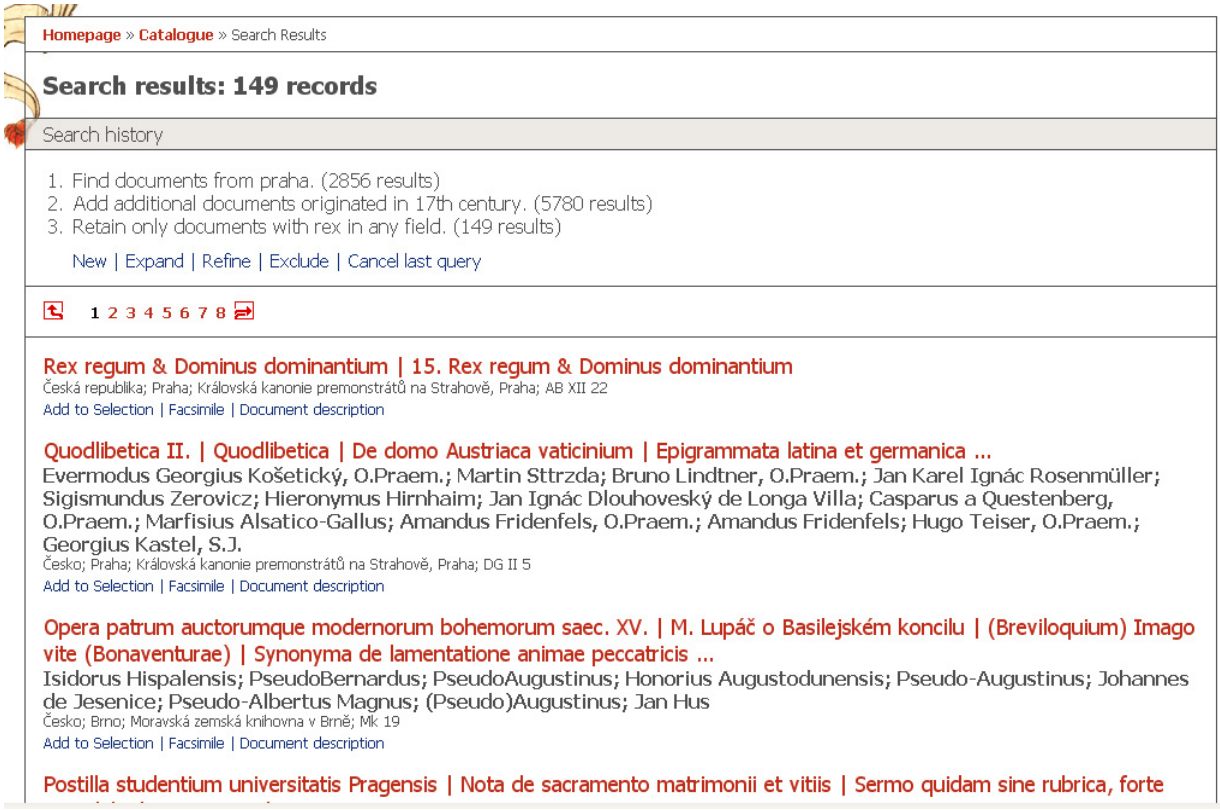

Fig. 19

Now again we continue to **exclude** from the search by clicking on the context link "Exclude" below the search history (fig. 20). We are interested to exclude all the documents from Brno.

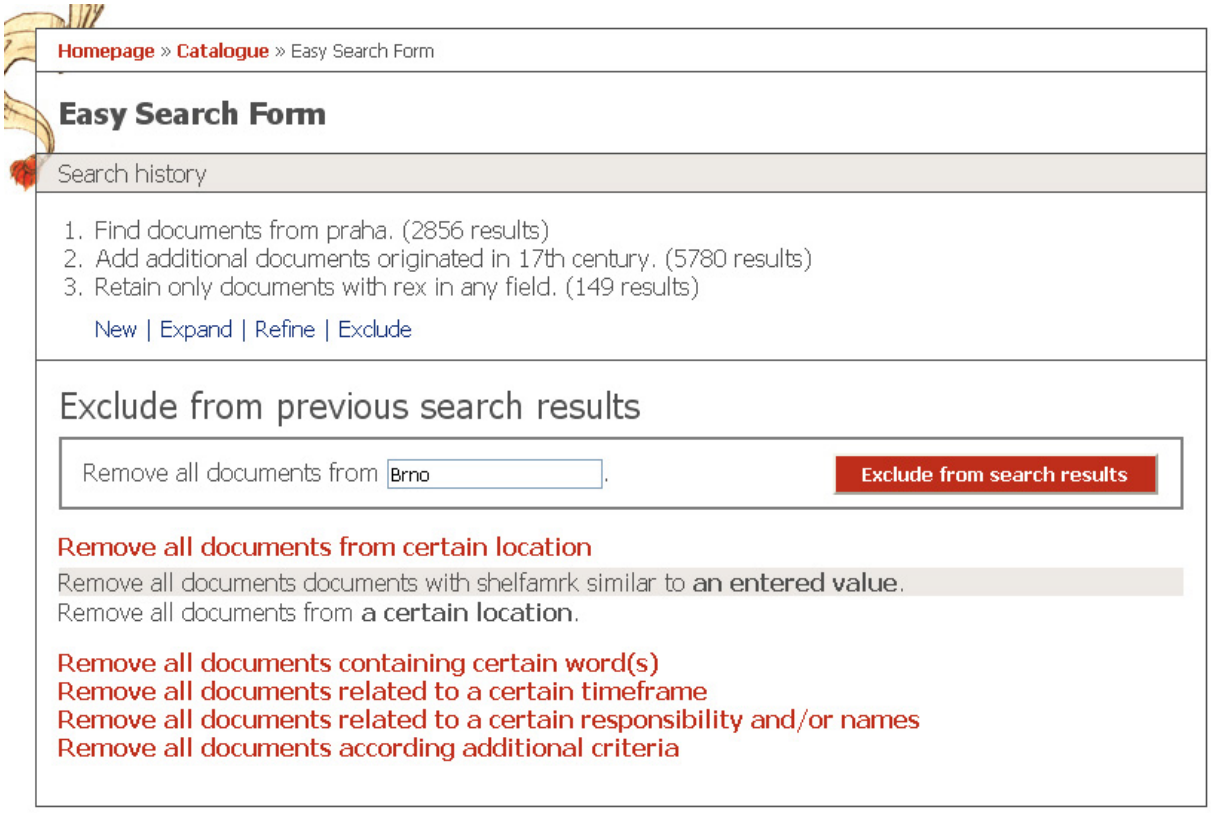

Fig. 20

The number of the records of this search appears on the top title Search results and the Search history shows at the point 1-2-3 the results of the previous searches and at the point 4 the result of the exclude search shows that the documents without the records from Brno are 130. (fig.21).

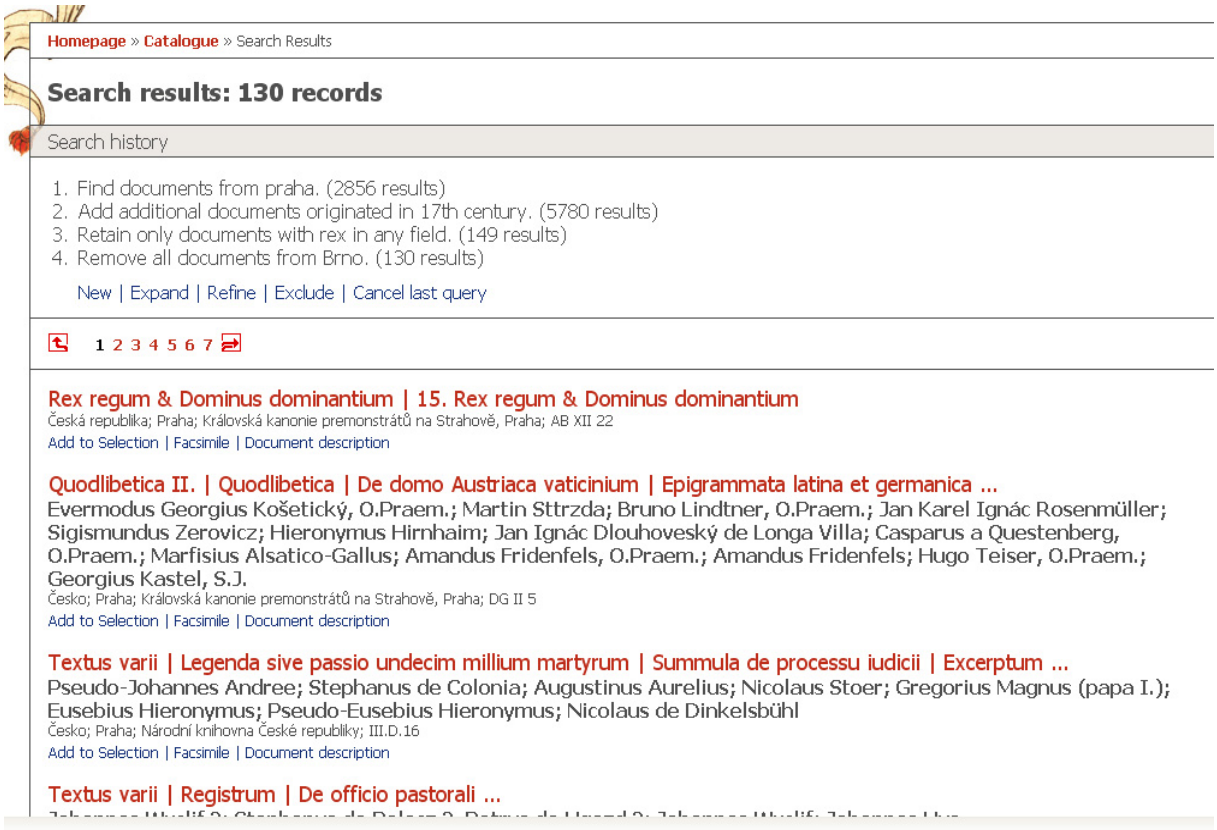

Fig. 21

Finally, every time we can **cancel last query** clicking on the context link "Cancel last query" below the search history (fig. 22) and to return to the previous search results.

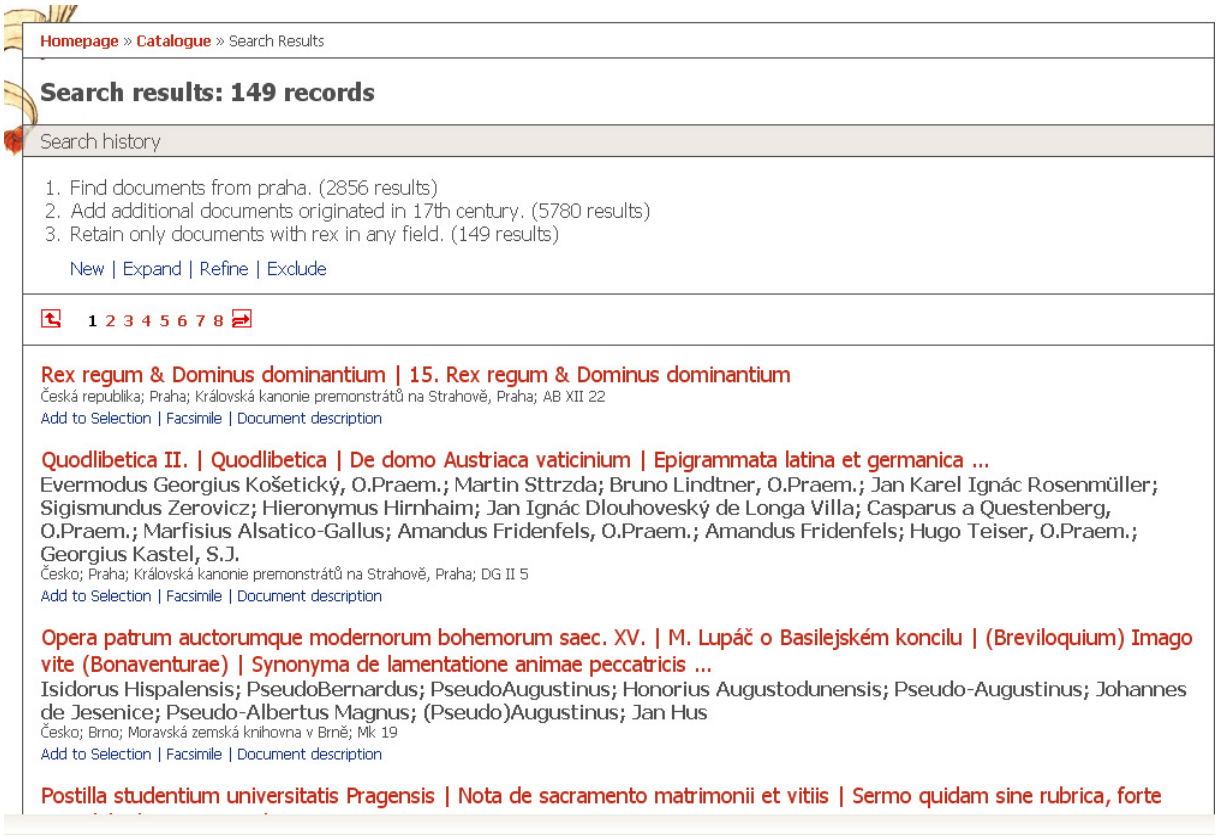

Fig. 22

4.5 Queries behind

The following query operators are now implemented into the search engine of Manuscriptorium. Some of them are the conventional some others are more specific. All of these are used behind the queries to enable better focused search results.

Operator	Description
AND &	standard operator AND
OR |	standard operator OR
WITHOUT -	standard operator WITHOUT
NOT	negation of a following part of query
>	greater than (for numbers, dates, text)
<	lower than (for numbers, dates, text)
>=	greater than or equal to (for numbers, dates, text)
<=	lower than or equal to (for numbers, dates, text)
a..b	range between a and b (a number/date/text greater than or equal to a and lower than or equal to b)
[a b n]	operands a and b have to be close to each other, their maximum distance is lower than or equal to n; the order of operands is important alternate expression: a___b where the number of _ represents the maximum distance of operands a/b/c is interpreted as /a b 1/ & /b c 1/
/a b n/	operands a and b have to be close to each other, their maximum distance is lower than or equal to n; the order of operands is not important alternate expression: a///b where the number of / represents the maximum distance of operands a_b_c is interpreted [a b 1] & [b c 1]
?	wildcard ?; multiple ? can be used to replace uncertain characters in an operand
*	wildcard *; can be used before, after or in a middle of a word
()	brackets enable construction of more complicated queries
;[space]	indicates a new query
;B	returns to a previous query
;Z	empties previous query results
;F	returns all database records
;O	the query will extend previous query results (OR)
;&	refines previous query results (AND)
;W	the query results are excluded form previous query results (WITHOUT)
%n	number of query filed to search in

As you can see above there are very specific operators which we call “intelligent” operators. Use of these can increase significantly the efficiency of queries but sometimes their usage can be difficult to use for certain groups of less computer-educated end-users. Their implementation into the new way of searching is therefore more than reasonable.

Apart form the operators above there are further options that affect the query results, these are:

diacritics tolerance	tolerance of diacritics – it is possible to set the query to fully or partially ignore diacritics (this may be usable when there are specifics characters with diacritics missing in the end-users keyboard)
alternates	it is possible to set the query to include usual graphemes into the tolerance, the system uses predefined tables that describe equal ways of writing of words, e.g. “'AE=OE=E” or “CRIST=CHRIST=XPT” (this is useful especially in the historical resources area)
shelfmark mode	it is possible to set the query in such a way that all characters that are not numbers or letters are ignored (both in the query and in the searched field(s)); this is extremely helpful when searching by shelf-marks

In the conventional searching these two options can be switched on/off using checkboxes. In the new way of searching the settings is implemented behind the query sentences depending on which type of information is retrieved.

4.6 Examples of predefined queries

To keep the interface easy to use we do not allow the user to use the various search options and search convention setting - the particular search options settings are fixed as a part of a particular query behind the sentences.

The following examples are taken from the user interface and demonstrate how the queries are formulated. Additionally it is demonstrated which parts of the records are searched by the queries (XPath expressions as applied above the TEI P5 ENRICH data are demonstrated too).

4.6.1 Example No. 1: Search for documents by their location

Find documents with shelfmark similar to O-C VIII 20.

Users input: “O-C VIII 20”
Searched query string: “O-C VIII 20”
Searched filed: “%x Shelf-mark, Identifier” (where x is appropriate field number)
Group of words: ON
Order of words: SIGNIFICANT, Distance: 0
Tolerance: OFF
Diacritical tolerance: OFF
Alternatives: OFF
* Auto-Complete: OFF

Filtering of insignificant characters: ON

The setting above reflects the facts that when searching for a particular Shelf-mark or ID number of a document, the shelf-mark/ID itself is known to the users. The last setting – filtering of insignificant characters – helps to receive results when the exact shape is uncertain – in our example the user knew the numbers and letters sequence but he/she was unsure about the other characters. The user entered "O-C VIII 20" and after the query and records were correctly interpreted the required document was retrieved, where the exact signature included a "/" instead of the "-" character and also one additional space character – see "O/ C VIII 20".

By applying the query above the following structures of source TEI P5 ENRICH schema XML records will be searched:

sourceDesc/msDesc/msIdentifier/descendant::idno
sourceDesc/msDesc/descendant::msPart/idno

4.6.2 Example No. 2: Search for documents by their location II

Find documents from Heidelberg.

Users input: "Heidelberg"
Searched query string: "%x Heidelberg" (where x is appropriate field number)
Searched fields: "Country", "Settlement", "Library"
Group of words: ON
Order of words: NOT SIGNIFICANT, Distance: OFF
Diacritical tolerance: OFF
Tolerance: OFF
Alternatives: OFF
* Auto-Complete: ON
Filtering of insignificant characters: OFF

You can see that multiple search fields are searched. The scope of TEI P5 ENRICH schema XML records that will be searched is:

sourceDesc/msDesc/msIdentifier/country
sourceDesc/msDesc/msIdentifier/settlement
sourceDesc/msDesc/msIdentifier/repository
sourceDesc/msDesc/msIdentifier/institution

4.6.3 Example No. 3: Search for documents related to a certain timeframe

Find documents originated in 14th century.

Users input: "14th century"
Searched query string: "%x >1500 & < =1400" (where x is appropriate field number)
Searched fields: "Year of Origin"
Group of words: OFF, Distance: OFF

Order of words: OFF
Diacritical tolerance: OFF
Tolerance: OFF
Alternatives: OFF
* Auto-Complete: OFF
Filtering of insignificant characters: OFF

The scope of TEI P5 ENRICH schema XML records that will be searched is:

sourceDesc/msDesc/head/origDate
sourceDesc/msDesc/history/origin/descendant::origDate
sourceDesc/msDesc/history/origin/descendant::date
sourceDesc/msDesc/descendant::msPart/head/descendant::origDate
sourceDesc/msDesc/descendant::msPart/history/origin/descendant::origDate
sourceDesc/msDesc/descendant::msPart/history/origin/descendant::date
sourceDesc/msDesc/msContents/descendant::msItem/docImprint/descendant::date
sourceDesc/msDesc/msContents/descendant::msItem/docImprint/descendant::origDate
sourceDesc/msDesc/descendant::msPart/msContents/descendant::msItem/docImprint/descendant::date
sourceDesc/msDesc/msContents/descendant::msItem/docImprint/descendant::origDate

4.6.4 Example No. 4: Search for documents containing certain word(s)

Find documents with sermones jacobi in the texts quoted from documents.

Users input: “sermones jacobi”
Searched query string: “%x sermones jacobi” (where x is appropriate field number)
Searched fields: “Text quoted from original” (=”Titles (All)”, “Rubrics”, “Incipits”, “Explicits”, “Colophons”, “Other quotations”)
Group of words: ON
Order of words: INSIGNIFICANT, Distance: max 5 words
Diacritical tolerance: ON
Tolerance: ON
Alternatives: ON
* Auto-Complete: ON
Filtering of insignificant characters: OFF

The scope of TEI P5 ENRICH schema XML records that will be searched is:

sourceDesc/msDesc/msContents/descendant::msItem/title
sourceDesc/msDesc/msContents/descendant::msItem/docTitle
sourceDesc/msDesc/descendant::msPart/sourceDesc/msDesc/msContents/descendant::msItem/title
sourceDesc/msDesc/descendant::msPart/msContents/descendant::msItem/docTitle
sourceDesc/msDesc/msIdentifier/msName
titleStmt/title
sourceDesc/msDesc/head/title

sourceDesc/msDesc/msContents/descendant::msItem/rubric
sourceDesc/msDesc/msContents/descendant::msPart/descendant::msItem/rubric
sourceDesc/msDesc/msContents/descendant::msItem/incipit
sourceDesc/msDesc/msContents/descendant::msPart/descendant::msItem/incipit
sourceDesc/msDesc/msContents/descendant::msItem/explicit
sourceDesc/msDesc/msContents/descendant::msPart/descendant::msItem/explicit
sourceDesc/msDesc/msContents/descendant::msItem/colophon
sourceDesc/msDesc/msContents/descendant::msPart/descendant::msItem/colophon
sourceDesc/msDesc/msContents/descendant::msItem/quote
sourceDesc/msDesc/msContents/descendant::msPart/descendant::msItem/quote

4.6.5 Example No. 5: Search considering responsibilities or names related to documents

Search for documents printed by Bartolomeus.

Users input: “Bartolomeus”
Searched query string: “(%x sermones jacobi) & (%y prints)” (where x and y are appropriate field number)
Searched fields: “Printer, Publisher” (x), “MNS document collections” (y)
Group of words: OFF
Order of words: INSIGNIFICANT, Distance: max 5 words
Diacritical tolerance: ON
Tolerance: ON
Alternatives: ON
* Auto-Complete: ON
Filtering of insignificant characters: OFF

The query above searches all prints where a name responsible for the print is Bartholomeus or Bartholomeaus ... etc.

The scope of TEI P5 ENRICH schema XML records that will be searched is:

sourceDesc/msDesc/msContents/descendant::name[../resp[@key='prt' or @key='pop' or @key='pbl']]
sourceDesc/msDesc/descendant::msPart/msContents/descendant::name[../resp[@key='prt' or @key='pop' or @key='pbl']]
sourceDesc/msDesc/msContents/descendant::persName[../resp[@key='prt' or @key='pop' or @key='pbl']]
sourceDesc/msDesc/descendant::msPart/msContents/descendant::persName[../resp[@key='prt' or @key='pop' or @key='pbl']]
sourceDesc/msDesc/msContents/descendant::msItem/docImprint/persName
sourceDesc/msDesc/msContents/descendant::msItem/docImprint/name[@type='person']
sourceDesc/msDesc/descendant::msPart/msContents/descendant::msItem/docImprint/persName
sourceDesc/msDesc/msContents/descendant::msItem/docImprint/name[@type='person']

5 Towards Deep Searching in Collections of Old Manuscripts by Extracting Semantic Information: proof-of-concept demonstrator.

ENRICH can provide seamless access to distributed knowledge on manuscripts. For that, advanced information retrieval methods comprising complex linguistic, cross-language and simple semantic operations on metadata have been implemented. On this basis, this contribution will discuss a simple use-case by introducing advanced semantic search facilities for metadata to enhance access to manuscripts. All ENRICH content partners agreed on providing knowledge on old manuscripts as TEI P5. And since the TEI provides – in addition to markup elements that are useful for describing manuscripts – means to record information about dates, people and places, the way for semantic processing has been cleared. A suite of software prototypes has been developed that implements a workflow to help prepare data which has been extracted from manuscripts for semantic browsing. The process of extracting information brought several obstacles to light that will be addressed. Possible solutions for these problems and relevant workings of Semantic Web research will be described.

The task description of ENRICH announces the provision of semantic searches that shall introduce an “intelligent operator”. But the notion of an “intelligent operator” still leaves a lot of room for interpretation and a simple use case referring to current Semantic Web research has been elaborated for demonstration purposes. (W3C 2009) One way to approach this problem is to reflect on how applying semantic operators can enhance a users' research experience beyond that of a simple full-text search. A historian who is pursuing research on the life of a specific historical person needs to acquire comprehensive knowledge about that person. Since names are notoriously spelled differently in old documents, a full-text approach will probably not be successful. Therefore, a system should be described that strives to provide semantic operators that are able to extract relevant bits of information from electronic manuscript descriptions.

Current Semantic Web research elaborated basic concepts and tools for information integration. In order to craft software components that implement the mentioned use-case, these developments should be exploited. In this regard, different aspects of Semantic Web research turned out to be useful. One of the most fundamental concepts in this area is the notion of a Uniform Resource Identifier (URI) that provides a way to globally and unambiguously identifies arbitrary material and immaterial things in the world. Furthermore, concepts like semantic markup and semantic triple stores have been exploited to facilitate semantic searches on ENRICH metadata. (Aduna 2009) To make use of these tools, certain information needs to be extracted from the ENRICH manuscript information.

A large amount of information in the humanities is derived from textual material. But even if texts have a clear structure and follow certain strains of arguments, from the perspective of automatic information processing they appear to be unstructured. In the context of ENRICH, all content providers agreed on providing information about old manuscripts as TEI P5 that comes with a certain predefined structure. (TEI 2009) First experiments showed that information about people, places and bibliographic entities could be extracted with reasonable effort. To support semantic searches that emancipate from simple field based evaluation

strategies, the extracted information has been mapped to a common structured vocabulary, the CIDOC CRM. (Dörr 2003) We have decided for the CRM because it provides the needed structural elements to establish semantic interoperability in the cultural heritage area. The respective parts of TEI that deal more exhaustive with names, dates, people and places have been mapped to the CRM. (Eide & Emil-Ore 2006) Figure 23 shows the basic structure of the CIDOC CRM.

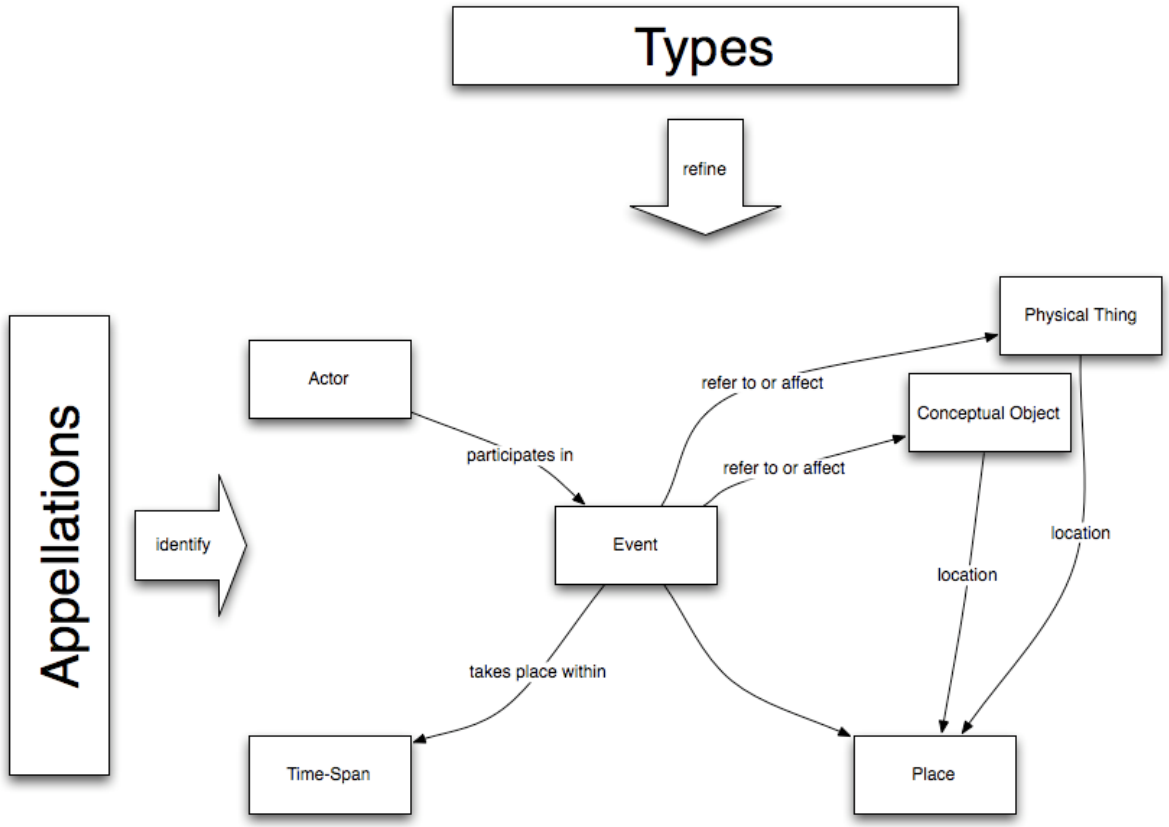

Fig. 23: The basic structure of the CIDOC CRM.[1]

The notion of Linked Data has become quite popular in the area of Semantic Web research, aiming at explicitly linking related information to achieve better knowledge discovery. (Christian Bizer et al. 2009) In this context, one area of problems has been identified that inhibits proper semantic processing of knowledge called "object matching" or "entity resolution". Historians for example are used to find references to historical people to be treated extremely inconsistent in old sources. Although, resolving these references is part of their day-to-day work, this task is laborious and extremely cost-intensive. (Eide 2008) Consequently, names that have been extracted from TEI documents do appear notoriously different although they are referring to the same person. Resolving these references automatically could lead to unintentional results because there is no authority that is accountable for each matching decision. A semi-automatic approach seems to be the most viable approach. Therefore, the demonstrator provides an environment that helps with

[1] The image follows http://cidoc.ics.forth.gr/docs/crm_for_gothenburg.ppt

resolving the extracted name information. It makes use of simple data mining techniques to fuel a recommender engine. (Elmagarmid et al. 2007) Fig. 14 shows a simple sandbox that helps users to manage co-reference information.

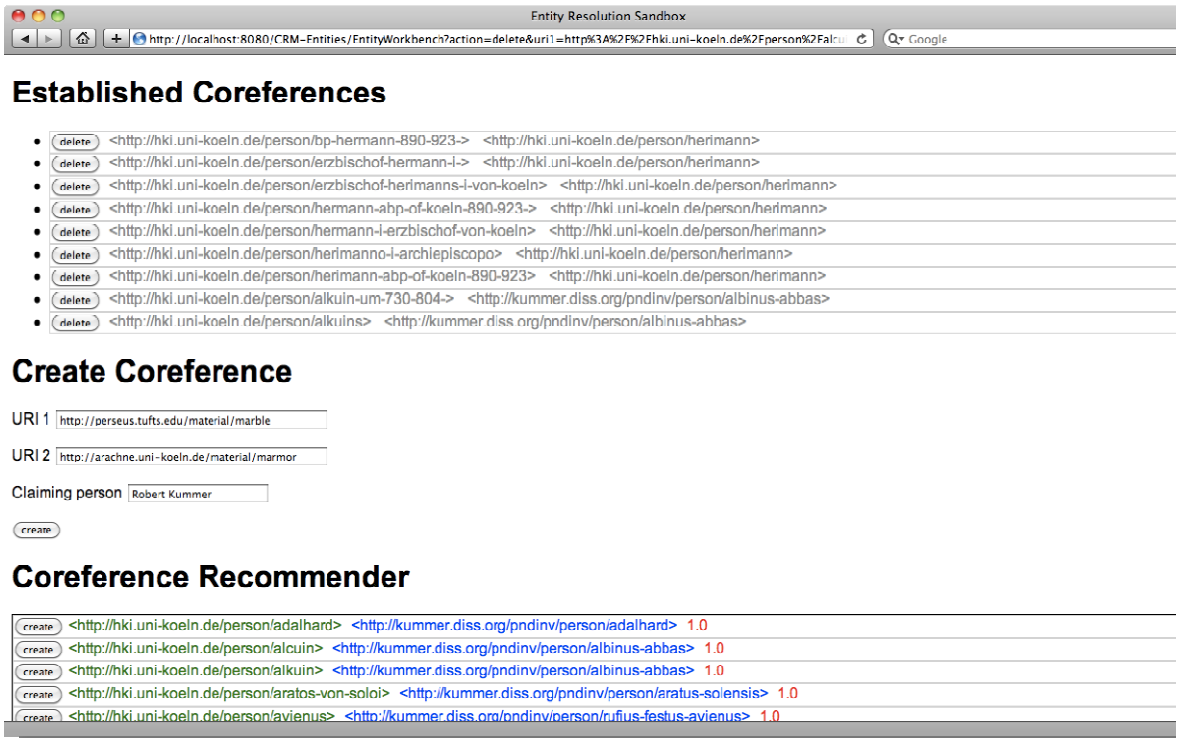

Fig. 14. A simple sandbox to manage coreference information.

The performance of such co-reference recommender can be improved by exploiting information that has already been structured in a certain way. Authority control for example has been traditionally cultivated in library and information science where it is an integral part of bibliographic control. (Sieglerschmidt 2007) Authority lists help disambiguating items that share the same heading, and collocating material that belongs together but appears to be different. Thus, authority lists inherently document information about the aforementioned co-references. However, while traditional libraries have been good at curating these files, no human being will be in the position to fulfill this task on a larger scale with growing amounts of digitally enriched material. In the area of Semantic Web research, one developing standard for organizing knowledge stands out: SKOS intends to provide a more straightforward approach to publish multilingual structured vocabularies. (Isaac & Summers 2008) Initiatives like “museumsvokabular” (Stefan Rohde- Enslin 2006) publish their vocabularies as SKOS. This should be exploited in the course of work on information integration.

A number of functional requirements have been collected so far that project a future system to support semantic operators in the scope of ENRICH. Demonstrating the thoughts that have been elaborated so far, various software components have been developed that support a continuous workflow, beginning with information extraction and ending with visualization of the results. Figure 25 shows the resulting information in a Semantic Web browser. The presented prototype produces data that can be used for advanced semantic searching in

Manuscriptorium but also in other aggregating projects that will publish this data like Europeana.

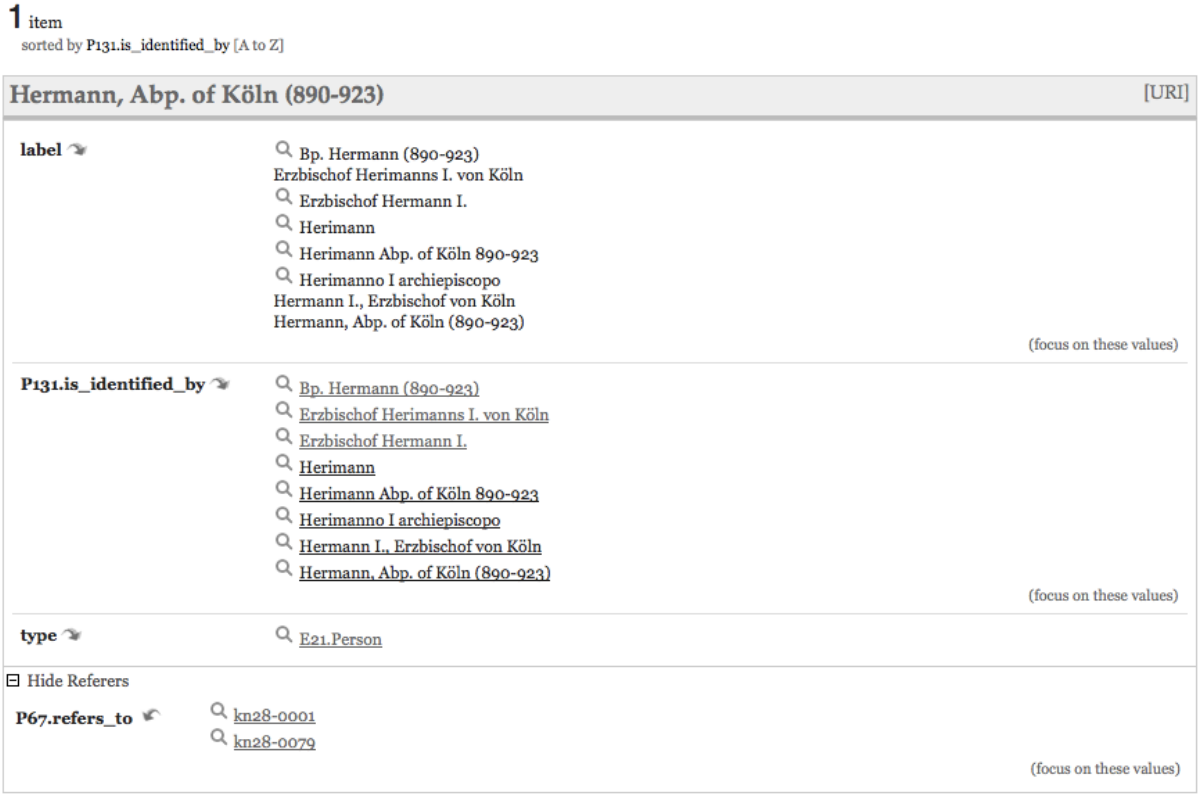

Fig. 25. Extracted names of persons have been loaded into a Semantic Web browser.

Furthermore, this semantic data can be stored in the TEI documents themselves. To this end, additional data – be it CIDOC CRM RDF or simple URIs – is embedded into <note>-tags in the TEI. This concept is called - concept as it doesn't necessarily have to be a TEI document, but rather any format versatile enough to contain other formats - a **Self-contained Object.**

A self-contained object is defined as an object in which all symbolic links are resolved internally or in its immediate dependencies – in this case, the object shall have all the data necessary to contextualize it in a given domain of intelligence.

In databases, most textual information is meta-information describing the real-world objects shown on the pictures. In libraries, most textual information is the actual content of the object. These two are to be incorporated in one single, self-contained object - sometimes pointers to URIs, sometimes as actual data. This would enable a researcher to review the primary sources (such as pictures, meta-data, texts and the like) on which an opinion in an, say, essay, is based without the need to do extensive research. Research which in the worst case would make travel to the library or the photographic archive in which the primary sources are hosted necessary, or, in a slightly less worse case, frustrating web searches for data long gone or only stored on project computers of long abandoned projects.

Take for example a dissertation on the subject of art history, and a database containing recent photos and metadata on the objects described in this dissertation. This metadata can now be incorporated into the TEI document (Figure 26) and be handled either as one complete entity (the book) or a collection of semantically interdependent data (the idea of the author of the book). Different scholars now have the possibility to either work with the whole book or harvest the metadata with or without the context.

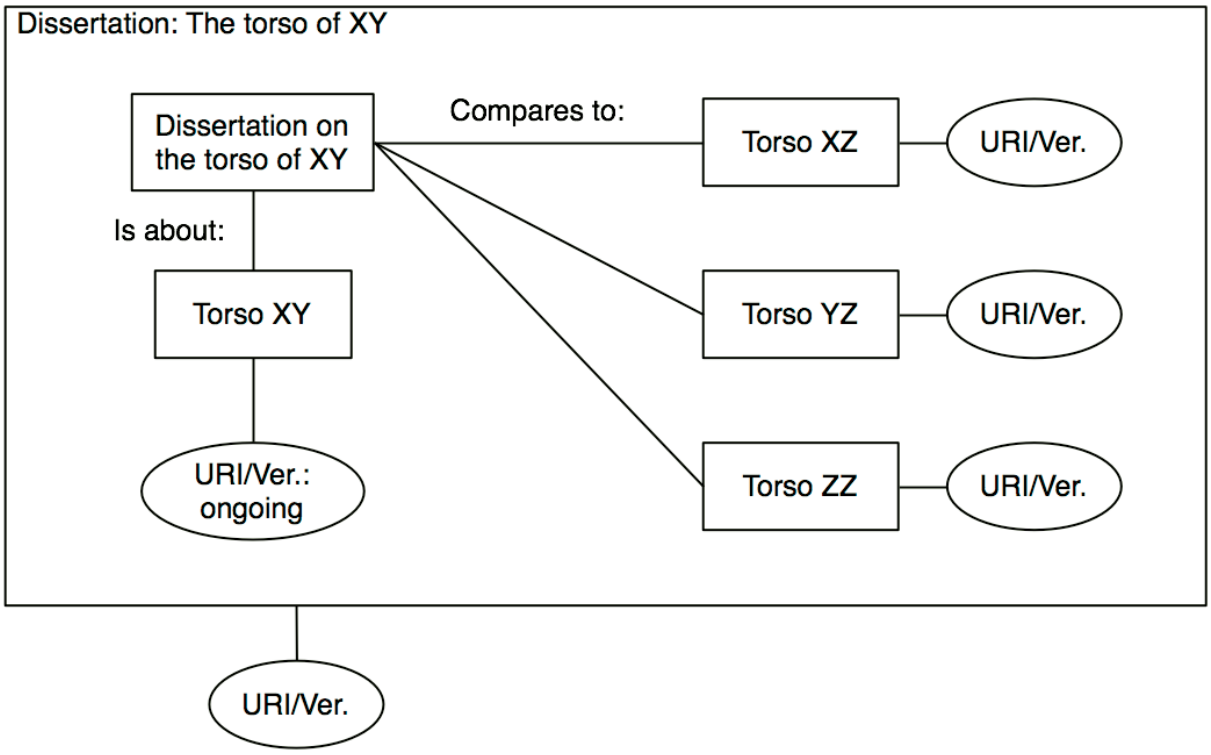

Fig. 26. Example of a selft contained object.

These objects can now be versioned to keep different opinions on the same text and follow the genesis of this self-contained object, and stored as a single, human-readable file, which can be versioned, copied, distributed and re-versioned. They can also be served via OAI-PMH, and researchers can choose how much of the document they need – for example, if a geoscientist needs radiocarbon data, and this data is embedded in a TEI document, which reflects a book on archaeology, he can harvest only this data without the archaeological context.

In this way completely heterogeneous data can be integrated in one easy-to-handle object, and the aforementioned semantic search technologies can harvest and structure the data contained within.

6 References

- Aduna, 2009. openRDF.org: Home. *openrdf.org: Home*. Available at: http://openrdf.org/ [Accessed January 5, 2009].
- Christian Bizer, Tim Berners-Lee & Tom Heath, 2009. Linked Data - The Story So Far. In *International Journal on Semantic Web & Information Systems, Vol. 5, Issue 3, Pages 1-22, 2009.*
- Dörr, M., 2003. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Mag*, 24(3), 75-92.
- Eide, Ø., 2008. What is co-reference? Available at: http://cidoc.mediahost.org/co_reference_wg%28en%29%28E1%29.xml [Accessed September 20, 2009].
- Eide, Ø. & Emil-Ore, C., 2006. TEI, CIDOC-CRM and a Possible Interface Between the Two. In *Digital Humanities*. pp. 62-4.
- Elmagarmid, A., Ipeirotis, P. & Verykios, V., 2007. Duplicate Record Detection: A Survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1), 1-16.
- Isaac, A. & Summers, E., 2008. SKOS Simple Knowledge Organization System Primer. Available at: http://www.w3.org/TR/skos-primer/ [Accessed November 25, 2008].
- Norvig, P. & Russell, S., 2003. *Artificial Intelligence: A Modern Approach* 2nd ed., Prentice Hall International.
- Sieglerschmidt, J., 2007. Knowledge organization and multilingual vocabularies. Vortrag auf der Jahrestagung "Managing the global diversity of cultural information" des Comite International pour la Documentation (CIDOC) in Wien 20.-22. August 2007. Available at: http://opus.bsz-bw.de/swop/volltexte/2008/280/ [Accessed February 5, 2009].
- Stefan Rohde-Enslin, 2006. museumsvokabular.de. Available at: http://museum.zib.de/museumsvokabular/ [Accessed September 20, 2009].
- TEI, 2009. TEI: Text Encoding Initiative. Available at: http://www.tei-c.org/index.xml [Accessed September 21, 2009].
- W3C, 2009. W3C Semantic Web Activity. Available at: http://www.w3.org/2001/sw/ [Accessed May 29, 2009].