Supporting Multilingual Bibliographic Resource Discovery with Functional Requirements for Bibliographic Records

Editor(s): Dimitrios A. Koutsomitropoulos, University of Patras, Greece; Eero Hyvönen, Aalto University, Finland; and Theodore S. Papatheodorou, University of Patras, Greece
Solicited review(s): Kate Byrne, Edinburgh University, UK; Ray R. Larson, University of California, Berkeley, U.S.A.; two anonymous reviewers

Hugo Manguinhas[a,b], Nuno Freire[a,b,c], Jorge Machado[a,d] and José Borbinha [a,b]
[a] *Department of Computer Science and Engineering, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*
E-mail: {hugo.manguinhas, nuno.freire, jlb}@ist.utl.pt
[b] *Instituto de Engenharia de Sistemas e Computadores, Rua Alves Redol, nº 9, 1000-029 Lisboa, Portugal*
[c] *The European Library, The National Library of the Netherlands, Willem-Alexanderhof 5*
2509 LK The Hague, Netherlands
[d] *Instituto Politécnico de Portalegre, Praça do Município, 7301-901, Portalegre, Portugal*
E-mail: jmachado@estgp.pt

Abstract. This paper describes an experiment exploring the hypothesis that innovative application of the Functional Requirements for Bibliographic Records (FRBR) principles can complement traditional bibliographic resource discovery systems in order to improve the user experience. A specialized service was implemented that, when given a plain list of results from a regular online catalogue, was able to process, enrich and present that list in a more relevant way for the user. This service preprocesses the records of a traditional online catalogue in order to build a semantic structure following the FRBR model. The service also explores web search features that have been revolutionizing the way users conceptualize resource discovery, such as relevance ranking and metasearching. This work was developed in the context of the TELPlus project. We processed nearly one hundred thousand bibliographic and authority records, in multiple languages, and originating from twelve European national libraries. This paper describes the architecture of the service and the main challenges faced, especially concerning the extraction and linking of the relevant FRBR entities from the bibliographic metadata produced by the libraries. The service was evaluated by end users, who filled out a questionnaire after using a traditional online catalogue and the new service, both with the same bibliographic collection. The analysis of the results supports the hypothesis that FRBR can be implemented for resource discovery in a non-intrusive way, reusing the data of any existing traditional bibliographic system.

Keywords: Resource Discovery, Multilingual Metadata, FRBR, MARC, UNIMARC

1. Introduction

Users often find traditional libraries' Online Public Access Catalogues (OPAC) display of results to be poor and inefficient, mainly due to the catalogues' usually linear structure and consequent multiple hits displayed on search results for equivalent resources [1]. This is especially true for long lists of multiple occurrences of a same work, when different manifestations are represented in different formats.

Web search engines like Google, and popular e-commerce interfaces such as Amazon, provide simple but powerful displays, which have been revolutionizing the way users conceptualize resource discovery. These systems provide three key features that distinguish them from most of the traditional OPAC interfaces: searching by propagation of a single query (**single search)**; reordering of search output according to relevance criteria (**relevance-ranked results**); and feedback to the user of similar terms ('did you mean') that may provide more relevant results than the current query (**relevance feedback**) [2]. Recently, Internet search engines also have been exploring the clustering of results according to their source (presenting the list of results through facets).

The FRBR (**Functional Requirements for Bibliographic Records)** [3] proposes a new conceptual model to serve as the basis for relating specific attributes and relationships to the various tasks that users perform when using bibliographic records. Rather than just changing cataloguing rules, the conceptual model introduced by FRBR has the potential to improve not only the intrinsic quality of the bibliographic information, but also the user experience when searching and browsing bibliographic catalogues.

The publication of the FRBR report started active discussions and created interest for many theoretical and research activities. An important constraint for the application of this model has been the difficulty in adapting the traditional cataloguing rules, and accordingly the existing cataloguing systems, to it. The complexity of the paradigm change is a real challenge for cataloguing professionals, thus still making it a high risk for software providers to consider it for their products. However, an important line of work has been the development of techniques for transforming legacy data into an FRBR implementation model, a process often referred to as **FRBRization** [4]. The millions of bibliographic records in use all over the world – created at a high cost – make the problem of FRBRization a very relevant one. Even though some library management systems are being designed from the start according to FRBR, they address the conversion of legacy data as a manual cataloguing process.

This paper presents an experiment aiming to take advantage of FRBR to integrate new features provided by web search technology into traditional library resource discovery systems. The main assumption behind this work is that it is possible to make good use of the semantically richer organization proposed by FRBR to: (i) **cluster** the results of a traditional search into groups of manifestations of the same work, and (ii) to **expand** each group with all the related FRBR manifestations and expressions by taking advantage of the linked information between the entities of the model. The second assumption is that one might be able to offer a better **ranking** of the results by taking advantage of the information contained within each FRBR work.

This work was developed in the context of the international project TELPlus[1], with the goal of exploring potential new techniques to improve the user searching experience with the service TEL (The European Library)[2]. In order to make it easier to assess the results, we decided to focus these experiments on a specific collection of works from Nobel Prize winners in Literature. These were found suitable due to the expected numerous editions and translations of these works into a large number of languages. Therefore, we collected as many as possible related MARC[3] [4] bibliographic and authority records, from libraries contributing to TEL.

The remainder of this paper presents an historical overview and the most relevant developments regarding FRBR, followed by an introduction to its conceptual and concrete models (ontologies). It continues with a description of our proposal, followed by a more detailed analysis of the process for the creation of FRBR data from the bibliographic records received from the libraries. Next, we provide an evaluation of the survey results. Finally, we discuss conclusions.

2. Historical Background and Related Work

This section presents the historical background and related work on improving the user experience with OPACs, particularly focusing on FRBR related experiments.

By the late 1980s, library professionals realized that great changes were happening in the library environment. The way information was being organized – especially considering the use of automated systems, new formats, electronic publishing, networked access and new web resources – needed careful reevaluation The Stockholm Seminar on Bibliographic Records, held in 1990 and sponsored by the IFLA Universal Bibliographic Control and International MARC (UBCIM) Programme and the IFLA Division of Bibliographic Control, was the moment chosen for a debate of these issues.

The participants in the Seminar were aware of the economic realities faced by libraries and the need to reduce the cost of cataloguing, but they also ac-

[1] http://www.theeuropeanlibrary.org/telplus/

[2] http://www.theeuropeanlibrary.org

[3] http://www.loc.gov/marc/

[4] http://www.ifla.org/unimarc

knowledged the importance of meeting the changing user needs. One of the nine resolutions approved in that Seminar led to a study of the functional requirements for bibliographic records (FRBR). The study's purpose was to delineate in clearly defined terms the functions performed by the bibliographic record with respect to various media, applications and user needs.

The study group was also charged with recommending a basic level of functionality and basic data requirements for records created by national bibliographic agencies. The result of that process was the creation of the IFLA Study Group on the Functional Requirements for Bibliographic Records, which produced the original report 'Functional Requirements for Bibliographic Requirements (FRBR)' [3]. This report describes a model that identifies and clearly defines the entities of interest to users of bibliographic records, the attributes of each entity, and the types of relationships that operate between entities. The intent was to produce a new conceptual model that would serve as the basis for relating specific attributes and relationships (reflected in the record as discrete data elements) to the various tasks that users perform when consulting bibliographic records.

The publication of the FRBR report started an active discussion and increased interest in many theoretical and research activities. The on-going discussions and outcomes can be followed in online resources such as the FRBR Bibliography[5], maintained by the FRBR Review Group, and the FRBR Blog[6].

One of these activities has been the development of a new generation of Integrated Library Systems (ILS) designed from the start according to the FRBR principles, including the OPAC interface. Examples of these systems are AustLit[7] (the Australian Literature Gateway), [5] a cooperative service involving eight universities and the National Library of Australia, and the Virtua ILS[8] from VTLS.

Another important line of work has been the development of processes to convert existing catalogues to new implementations following the FRBR model, which is often referred to as FRBRization. Some relevant studies show us that much of the information needed to FRBRize catalogues is already present in MARC data [6],[7]. The challenge of FRBRizing legacy data and the reality of current catalogue systems were also addressed in [8], which stresses that, 'to make the transition to FRBR possible, it is necessary to extract the FRBR structure from existing data'. Following this line of work, in [4] an attempt was made to analyse MARC records and determine what attributes could best be used for FRBRization.

Several efforts have been undertaken to develop algorithms for the FRBRization of bibliographic data, the OCLC FRBR Work-Set Algorithm[9] being the most important reference. This algorithm, used in both OCLC FictionFinder[10] and OCLC Curiouser[11] prototypes, is focused on the clustering of data looking at the FRBR group 1 (Work, Expression, Manifestation and Item) but with special attention paid to Works. Subsequent efforts tried to take FRBRization further by allowing clustering of all entities defined in the FRBR. An example is the BibSys FRBR conversion tool [9], which actually builds a FRBR-based structure from bibliographic records and tries to cluster all entities of that structure by comparing their keys (combination of properties found in a FRBR entity). In spite of all these efforts, authors further explain that 'algorithms for eliciting FRBR structure' will only work as well as the bibliographic data on which they are based [10].

On the other hand, web search engines like Google, and popular e-commerce websites such as Amazon, provide simple, powerful displays that have been revolutionizing the way users search for information.

One of the key features of these systems is the ability to converge searching into a **single search** form, which is expanded to all information resources. This simplifies the access to resources by 'guiding users to where they are most likely to find results quickly' and therefore 'should satisfy the needs of the majority of users' [2].

Another important technique is the ranking of search output according to its relevance (**relevance ranking**). 'This feature transformed the way people search for information. Before these, most search technology focused not on bringing relevant material to the top of the list, but on eliminating irrelevant material from the result set. This approach did not always make it easy to find material if the result set was large. It made it harder to search very large data-

[5] http://www.ifla.org/VII/s13/wgfrbr/bibliography.htm

[6] http://www.frbr.org/

[7] http://www.austlit.edu.au

[8] http://www.vtls.com/products/virtua

[9] http://www.oclc.org/research/projects/frbr/algorithm.htm

[10] http://fictionfinder.oclc.org/

[11] http://www.oclc.org/research/projects/curiouser/default.htm

bases, within which many items might be somewhat relevant.' [11].

Another technique is to offer feedback to the user about similar keywords that may provide more relevant results than the current query (**relevance feedback**). These alternative searches are typically found next to the search query and accompanied by 'more like this' or 'did you mean' expressions. Works like [12] show that this kind of feature can also be provided to users of library catalogues. Their work consisted of applying relevance feedback strategies to analyse the content of the records retrieved and identify terms that are likely to retrieve other relevant documents. The perceived assumption behind that seems to be the one that 'if a term occurs in most of the records found relevant by the user and occurs in few non-relevant records, then it is likely to retrieve other relevant records'.

In fact, techniques like relevance-ranking and relevance feedback were already present in some experimental catalogues [13], called next-generation OPACs, which date back to the late eighties and early nineties (long before the web-search engines). These next-generation OPACs took advantage of the research on information retrieval (IR) to add new features to catalogues. However, according to the work by [11], the standard relevance-ranking algorithms present in next-generation OPACs have been mostly developed for full-text documents and additionally attempted to improve results by taking advantage of the highly structured nature of bibliographic records (therefore, they rely on a set of principles for ranking data contained in specific fields in the bibliographic record).

Finally, another technique worth mentioning is the **clustering** of results according to some relevant criteria, as demonstrated by the OPAC of the Research Libraries Information Network (RLIN), which used similar techniques to cluster bibliographic records with the same title [14].

Although successful, these techniques only became popular after the emergence of web-search engines like Google. Sources like [2] point out that 'the popularity of the web appears to have influenced users' mental models and thus their expectations and behaviour when using a web-based OPAC interface'. The same source also 'attribute the increase to the prevalence of web search engines and suggest that metasearching, relevance-ranked results, and relevance feedback ("more like this") are now expected in user searching and should be integrated into online catalogues as search options'.

3. An Ontology for FRBR

The conceptual model as defined in the initial FRBR report is composed of ten entities divided into three groups. The first group (Endeavours) is comprised of the products of intellectual or artistic endeavours named or described in bibliographic records: **Work**, **Expression**, **Manifestation** and **Item**. The second group (Responsible Entities) is comprised of the entities responsible for the intellectual or artistic content, the physical production and dissemination or the custodianship of such products: **Person** and **Corporate Body**. The third group (Subjects) is comprised of an additional set of entities that serve as the subjects of intellectual or artistic endeavour: **Concept**, **Object**, **Event** and **Place**. The model also defines the attributes and relationships to be applied to each entity.

Figure 1 shows an overview of the base entities of the FRBR model and the most important relationships defined to relate them. Note that only a brief overview of the model is given in this section. For a more detailed description of the model, see Sections 4 and 5 of the FRBR report [3]. Note also that all the definitions presented in this section were obtained from the same report.

A **Work** is an abstract entity; there is no single material object one can point to as the work. The work is recognized through individual realizations or expressions of the work, but the work itself exists only in the commonality of content between and among the various expressions of the work. A work is defined by a title, a form (e.g. novel, play, poem, map or painting), a date and context of its creation, and an intended termination and audience for the work. Other attributes can be defined that are specific to musical (e.g. medium, key) and cartographic works (e.g. coordinates, equinox).

An **Expression** is the intellectual or artistic realization of a work in the form of alphanumeric, musical or choreographic notation; sound, image, object, movement etc.; or any combination of such forms. The boundaries of the entity Expression are defined, however, so as to exclude aspects of physical form, such as typeface and page layout, that are not integral to the intellectual or artistic realization of the work as such. Following this, an Expression is defined by one or more titles, a form (the means by which the Work is realized), a date of creation, a language, its extensibility and revisability, its extent (a quantification of its content, e.g., number of words in a text, images in a comic strip), a summarization of its content, its context (e.g. art period) and a description of the critical response and use restrictions. As with Works, Expressions can be further defined by attributes that are specific to a particular kind of Expression (e.g. serial, music notation, sound, cartographic, projected and remote sensing images).

A **Manifestation** is the physical embodiment of an expression of a work. It encompasses a wide range of materials, including manuscripts, books, periodicals, maps, posters, sound recordings, films, video recordings, CD-ROMs and multimedia kits. As an entity, manifestation represents all the physical objects that share the same characteristics, in respect to both intellectual content and physical form. As such, it is defined by an identifier, one or more titles, a statement of responsibility, an edition/issue designation, a publisher/distributor (modelled as an object property in this work), a place of publication and/or distribution, fabricator/manufacturer (also an object property), a series statement, a designation of the form, extent, medium and dimensions of the physical carrier, the source of acquisition and capturing mode, terms of availability and access restrictions. A manifestation can be further detailed through specific attributes for printed, hand-printed, serial, sound recording, image, microform, visual projection and electronic resource materials.

An **Item** is a concrete entity corresponding to a single exemplar of a manifestation. It is, in many instances, a single physical object (e.g., a copy of a one-volume monograph or a single audio cassette). There are instances, however, where the item comprises more than one physical object (e.g., a monograph issued as two separately bound volumes or a recording issued on three separate compact discs). An Item is defined by an identifier; a fingerprint; descriptions of its provenance, exhibition and treatment history; condition; marks/inscriptions and access restrictions.

A **Person** is an individual involved in the creation or realization of a Work (e.g., as author, composer, artist, translator, etc.), or who is the subject of a Work (e.g., as the subject of a biographical or autobiographical work, of a history, etc.). A Person is defined by a name, a date of birth and death (if applied), a title and an attribute for additional information.

A **Corporate Body** is an organization or group of individuals and/or organizations acting as a unit that are identified by a particular name, including occasional groups and groups that are constituted as meetings, conferences, congresses, etc. It is defined by a name, a number/identifier, a place and date associated with the Corporate Body and an attribute for additional information.

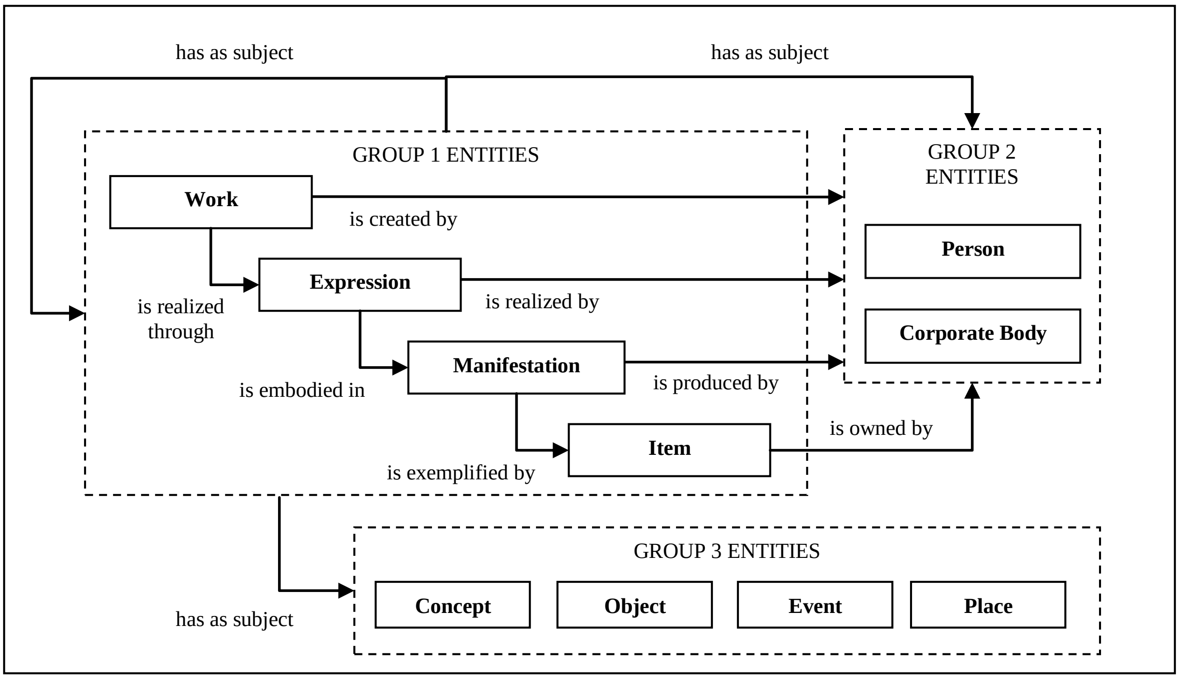


Fig. 1. The conceptual model defined in the FRBR report.

A **Concept** is an abstract notion or idea. An **Object** is a material thing, which may encompass both animate and inanimate objects occurring in nature; fixed, movable, and moving objects that are the product of human creation; or objects that no longer exist. An **Event** is an action or occurrence in time (e.g. historical events, epochs and periods of time). A **Place** is a location (terrestrial or extra-terrestrial) which might be expressed as a historical or contemporary name, a geographic feature, a geo-political jurisdiction, etc. All subjects (Concept, Object, Event or Place) are only defined by a term. However, some additional properties were defined to accommodate information specific to each kind of subject.

As mentioned in the FRBR report, 'The model operates at the conceptual level; it does not carry the analysis to the level that would be required for a fully developed model'. For the purpose of this work, therefore, a concrete specification of the model (an ontology as defined by Gruber in [15]) was needed, both to serve as the basis for the integration of several data sources and to expose the information for the semantically enriched search.

After a careful analysis of the two best-known ontologies that follow the FRBR conceptual model, the FRBR in RDF[12] ontology was chosen. It defines only the basic entities, keeping the data structure very simple, and offers a digital format for encoding the data. An alternative could have been the more complex FRBRoo [16] ontology (about 23 classes), which, although it is a well-formed and mature ontology, has no available formal representation.

In this study, the FRBR in RDF ontology was extended to include as class properties all of the attributes and relationships defined in the FRBR conceptual model. Appendix A gives an overview of the class properties that were actually used in this work. It also provides some statistical information on the number of classes and properties found for each class.

Although all properties defined in the FRBR were specified in the ontology, only a subset of them was actually used since they were not available in the source data or no structured or semi-structured field was used to encode them. As examples are some of the properties related to musical, cartographic and serial works, and some very specific properties like: intended termination and context of a work; revisability, critical response and user restrictions for expressions; and typeface, type size, foliation, collation, polarity, reduction ratio, generation for manifestations.

For the remainder of this paper, the term 'FRBR ontology' will be used to refer to the extended version of the FRBR in RDF ontology designed for this work.

4. Improving Resource Discovery

A number of studies suggest that with the prevalence of web search engines, users now expect to see features found in these systems when searching an OPAC [2]. On the other hand, the work developed under the FRBR has provided a semantically richer and more mature model for the representation of bibliographic information. The idea behind this work is to combine these two perspectives, taking advantage of their key features.

To accomplish this, a search interface was designed that offers the ability to present the results of a search in two alternative options: as a traditional OPAC, and using a semantically enriched search, which takes advantage of the FRBR. This ability to switch between both options enables the user to easily compare and evaluate the results. Searching in both modes can be done using a **simple search** or, as with a traditional OPAC, using multiple metadata elements. The semantically enhanced search is done the same way as for the traditional OPAC, but before presenting the results to the user, they are sent to a new service, outside the OPAC, where they are linked, clustered, expanded and reordered according to the FRBR.

This new service, called **semantic cluster**, is thus responsible for receiving a list of bibliographic references and delivering a tree-like structure containing the same references, but now clustered, expanded and reordered.

The **clustering of results** in the semantic cluster is done by taking advantage of the relationships (of the types realization, embodiment and exemplar) between the first group of entities (work, expression, manifestation and item). It is important to note that in most cases the bibliographic records are conceptually placed at the manifestation level of the FRBR conceptual model. The list of bibliographic references can be mapped into the corresponding manifestations, which can then be clustered according to their indirect relationship with a work. This interface seems closer to what Internet search engines do when grouping references from the same site or Internet domain.

[12] http://vocab.org/frbr/core.html

After the results are grouped into clusters, each cluster is **expanded** with all the manifestations available for the work. This is particularly important for reaching expressions of a work in languages other than the language of the query. Each cluster is **ranked** by combining the ranking given by the OPAC (which in this experiment uses common ranking algorithms used in IR) with the number of manifestations within the cluster. The idea is to take advantage of the default relevance ranking of the search engine and expand it with information about the number of manifestations of a work. It is important to note that this approach assumes the OPAC already applies standard ranking; therefore, the purpose of this work is to complement and not replace these algorithms. The formula logarithmically lowers the weight of the number of manifestations; otherwise the ranking given by the search engine would have very little influence on the final ranking. The logarithm of base ten was found adequate after analysing the ranking results delivered by the OPAC. Although it provides satisfactory results, a better evaluation of this function would be required. The complete formula for the ranking of cluster is shown in Eq. (1).

(1) Cluster ranking = 'Highest ranking manifestation' x log10 ('Total number of manifestations')

For the ranking of manifestations within each cluster, the formula uses first the publication date of the manifestation (from the lowest date – least recent – to the highest – most recent) and, if absent, uses the default ranking. This way the sorted results give more relevance to the first publications of a work.

The **relevance feedback** technique is used by both OPACs, which consists of finding the most relevant terms returned by the engine and using them to improve the search task. In the case of the FRBR OPAC, this function was adapted to re-rank clusters using score functions that take advantage of the cluster index in the semantic cluster. The main idea is to consider clusters as documents in score functions.

The semantic cluster uses a common repository fine-tuned for this purpose. To improve its performance, the schema of this repository was designed to hold only the essential information needed to cluster, expand and rank the results. Also, the indexes were designed to allow for a fast retrieval of the information. On average, the semantic cluster takes about a quarter of a second to do its work, which is essential for a good response time of less than one second.

Prior to being used for clustering, the repository must be loaded with the bibliographic information provided by the twelve participating libraries. Since the information encoded in the original MARC records is defined in a semantically poorer model, the records need to be transformed into the FRBR ontology, a process often called as FRBRization (see Section 5). This is a very time-consuming process, due to the amount of work required to normalize, extract and aggregate all the FRBR entities. For this reason, the work is done in a prior stage and not at runtime. Additionally, both new clustering algorithms and user feedback can further improve pre-built clusters.

The service interface of the semantic cluster was designed for easy integration with any common OPAC. This way any OPAC can take advantage of the new features provided by the application of FRBR, assuming that the semantic cluster is updated with the same records as the OPAC. The integration between the two services, the OPAC and the semantic cluster, is shown in Figure 2. Steps 1 to 4a show the interactions between the components of the OPAC during a search request using the traditional OPAC interface, while steps 1 to 8b show the interactions when the semantically enhanced search is used.

Figure 3 shows the presentation of the search results produced by a traditional OPAC. It shows, for the specific collection that was used, the fourteen results of a generic search for the title "The Outsider" (also commonly translated to English, especially in the United States, as "The Stranger'). Figure 4 shows the same results, but rearranged according to the FRBR. The results are grouped into six clusters (corresponding to six different works) containing 206 results (each one corresponds to a different manifestation and links to the original bibliographic record), and not just fourteen results. This is a result of the aggregation of all the manifestations associated with each work, which in the case of the first search result corresponding to the work ('L'Étranger') gathers about 151 results. Another interesting feature is the clustering of titles in different languages, which can be seen in the fourteen titles found for the first work ('L'Étranger').

Another user interface was also built to search entities found in the bibliographic records. The idea is the same as for the OPAC, but instead of searching for bibliographic information, the user may search for responsible entities (person, corporate body or family). The added value of this interface is the ability to access all the information related to a given authority: along with the names, dates, and roles of an authority, the user may also see all the identifiers

(Authority Record Number) that were given in each library and thus access its source record.

This interface also uses the semantic cluster to retrieve the list of responsible entities that match a given query. In this scenario, the traditional OPAC is not used, since it is designed for bibliographic retrieval. The query is handled directly by the semantic cluster through a specific service interface.

The **clustering** and **expansion** of results for authority information is already fulfilled by the repository, since the query is run over the cluster indexes and not the flat structure of the OPAC. Again, to improve performance, the cluster indexes use specialized text indexes over the name property of authors. When loading the repository, the information (defined in the FRBR ontology) must already be clustered. This is done in the entity aggregation step of the FRBRization process.

The **ranking** of each authority cluster is calculated as the sum of the ranking for each role relationship with the authority. This ranking is itself calculated by combining the ranking of the role relationship (a controlled list of roles and predefined rankings) and the number of endeavours related to that authority through that specific role. The main idea is to give more weight to an authority with more endeavours associated with it, while giving more relevance to some roles (e.g. author, creator) than others (e.g. translator, owner), which are likely to be less relevant to the user. The complete formula for the ranking of the cluster is shown in Eq. (2).

(2) Authority cluster ranking = Σ 'ranking of a given role' x log10 ('Total number of endeavour of a given role')

Both the authority and bibliographic search interfaces have links to each other. The user may switch between Group 1 entities (work, expression, manifestation and item) and Group 2 entities (person, corporate body and family) using linked-data relationships of the FRBR ontology. Figure 5 shows the author search results for the query 'Camus' which retrieved eleven results (corresponding to eleven different authorities) from the repository. One of them is for the Nobel Prize-winning author 'Albert Claude Camus'.

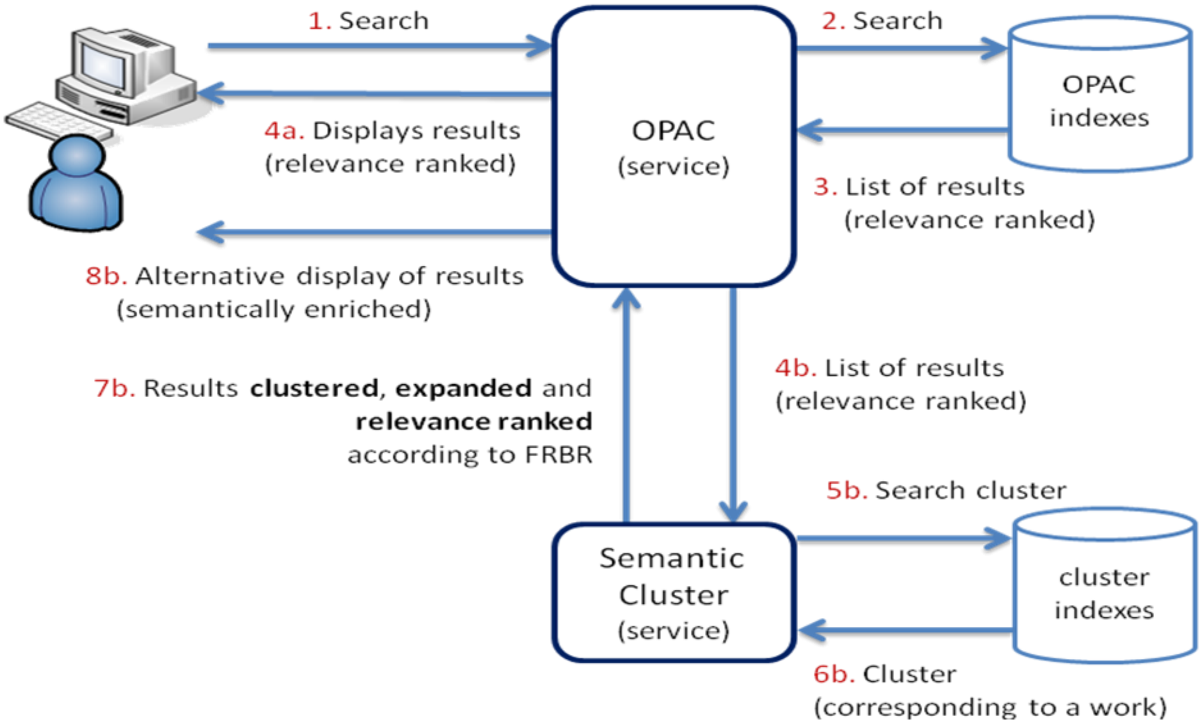
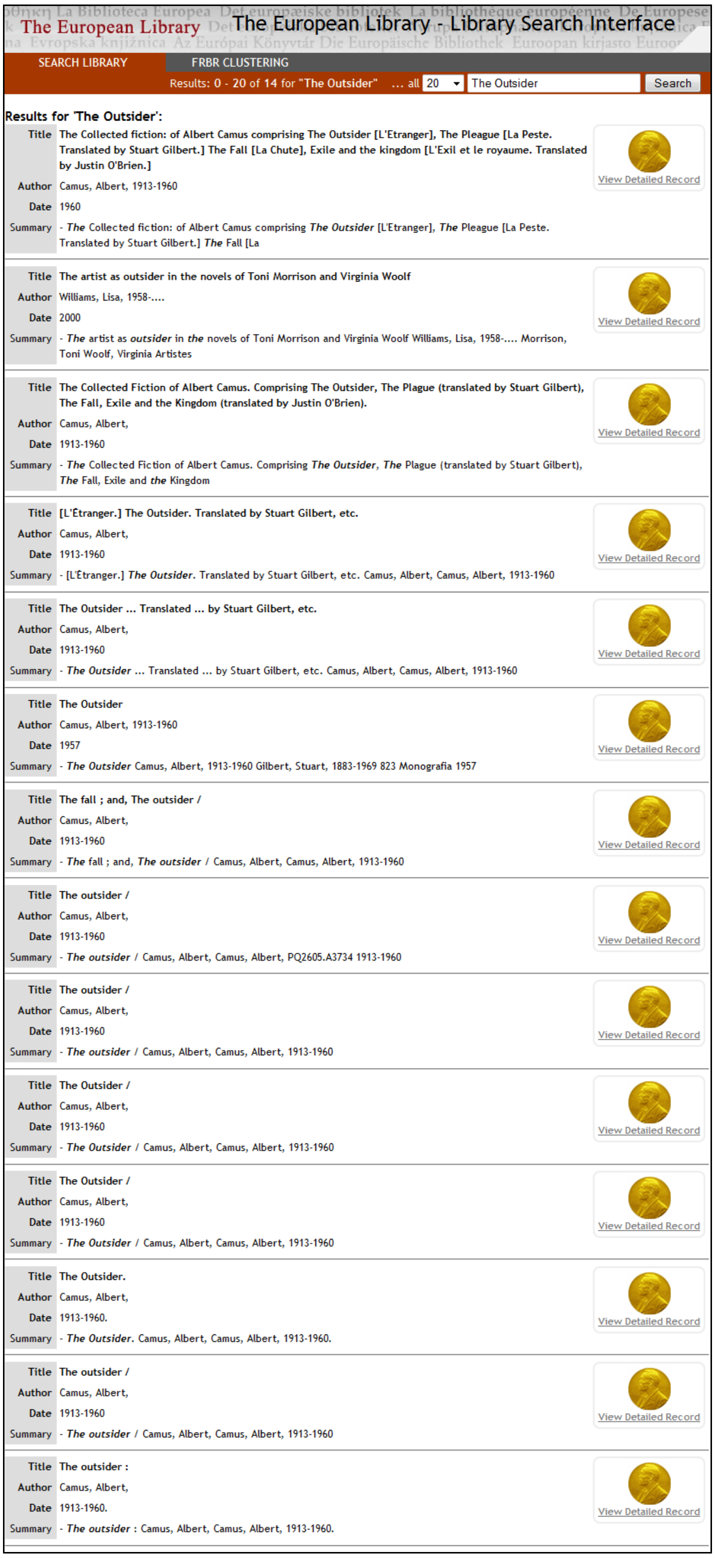

1. Search
2. Search
OPAC indexes
OPAC (service)
4a. Displays results (relevance ranked)
3. List of results (relevance ranked)
8b. Alternative display of results (semantically enriched)
7b. Results clustered, expanded and relevance ranked according to FRBR
4b. List of results (relevance ranked)
5b. Search cluster
Semantic Cluster (service)
cluster indexes
6b. Cluster (corresponding to a work)

Fig. 2. Overview of the interactions between the components of the OPAC for a search using both traditional (steps 1 to 4a) and semantically enhanced (steps 1 to 8b) search interfaces.


The European Library
The European Library - Library Search Interface
SEARCH LIBRARY
FRBR CLUSTERING
Results: 0 - 20 of 14 for "The Outsider"
... all
20
The Outsider
Search

Results for 'The Outsider':

Title The Collected fiction: of Albert Camus comprising The Outsider [L'Etranger], The Pleague [La Peste. Translated by Stuart Gilbert.] The Fall [La Chute], Exile and the kingdom [L'Exil et le royaume. Translated by Justin O'Brien.]
Author Camus, Albert, 1913-1960
Date 1960
Summary - ***The*** Collected fiction: of Albert Camus comprising ***The Outsider*** [L'Etranger], ***The*** Pleague [La Peste. Translated by Stuart Gilbert.] ***The*** Fall [La
View Detailed Record

Title The artist as outsider in the novels of Toni Morrison and Virginia Woolf
Author Williams, Lisa, 1958-....
Date 2000
Summary - ***The*** artist as ***outsider*** in ***the*** novels of Toni Morrison and Virginia Woolf Williams, Lisa, 1958-.... Morrison, Toni Woolf, Virginia Artistes
View Detailed Record

Title The Collected Fiction of Albert Camus. Comprising The Outsider, The Plague (translated by Stuart Gilbert), The Fall, Exile and the Kingdom (translated by Justin O'Brien).
Author Camus, Albert,
Date 1913-1960
Summary - ***The*** Collected Fiction of Albert Camus. Comprising ***The Outsider***, ***The*** Plague (translated by Stuart Gilbert), ***The*** Fall, Exile and ***the*** Kingdom
View Detailed Record

Title [L'Étranger.] The Outsider. Translated by Stuart Gilbert, etc.
Author Camus, Albert,
Date 1913-1960
Summary - [L'Étranger.] ***The Outsider***. Translated by Stuart Gilbert, etc. Camus, Albert, Camus, Albert, 1913-1960
View Detailed Record

Title The Outsider ... Translated ... by Stuart Gilbert, etc.
Author Camus, Albert,
Date 1913-1960
Summary - ***The Outsider*** ... Translated ... by Stuart Gilbert, etc. Camus, Albert, Camus, Albert, 1913-1960
View Detailed Record

Title The Outsider
Author Camus, Albert, 1913-1960
Date 1957
Summary - ***The Outsider*** Camus, Albert, 1913-1960 Gilbert, Stuart, 1883-1969 823 Monografia 1957
View Detailed Record

Title The fall ; and, The outsider /
Author Camus, Albert,
Date 1913-1960
Summary - ***The*** fall ; and, ***The outsider*** / Camus, Albert, Camus, Albert, 1913-1960
View Detailed Record

Title The outsider /
Author Camus, Albert,
Date 1913-1960
Summary - ***The outsider*** / Camus, Albert, Camus, Albert, PQ2605.A3734 1913-1960
View Detailed Record

Title The outsider /
Author Camus, Albert,
Date 1913-1960
Summary - ***The outsider*** / Camus, Albert, Camus, Albert, 1913-1960
View Detailed Record

Title The Outsider /
Author Camus, Albert,
Date 1913-1960
Summary - ***The Outsider*** / Camus, Albert, Camus, Albert, 1913-1960
View Detailed Record

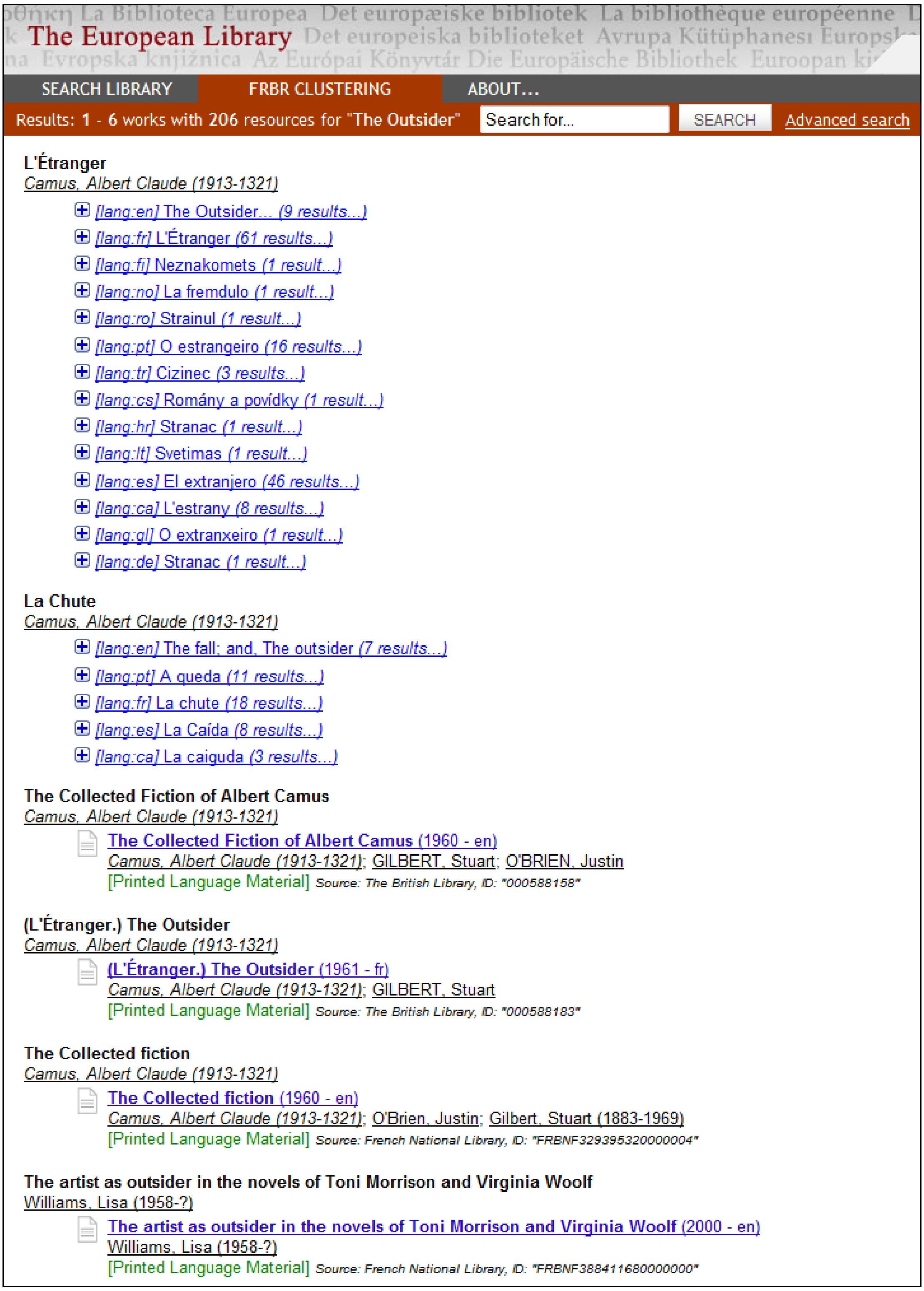
Title The Outsider /
Author Camus, Albert,
Date 1913-1960
Summary - ***The Outsider*** / Camus, Albert, Camus, Albert, 1913-1960
View Detailed Record

Title The Outsider.
Author Camus, Albert,
Date 1913-1960.
Summary - ***The Outsider***. Camus, Albert, Camus, Albert, 1913-1960.
View Detailed Record

Title The outsider /
Author Camus, Albert,
Date 1913-1960
Summary - ***The outsider*** / Camus, Albert, Camus, Albert, 1913-1960
View Detailed Record

Title The outsider :
Author Camus, Albert,
Date 1913-1960.
Summary - ***The outsider*** : Camus, Albert, Camus, Albert, 1913-1960.
View Detailed Record

Fig. 3. Search results for ‘The Outsider’ using the traditional OPAC interface.



Results: 1 - 6 works with 206 resources for "The Outsider"

L'Étranger
Camus, Albert Claude (1913-1321)

- ⊞ [lang:en] The Outsider... (9 results...)
- ⊞ [lang:fr] L'Étranger (61 results...)
- ⊞ [lang:fi] Neznakomets (1 result...)
- ⊞ [lang:no] La fremdulo (1 result...)
- ⊞ [lang:ro] Strainul (1 result...)
- ⊞ [lang:pt] O estrangeiro (16 results...)
- ⊞ [lang:tr] Cizinec (3 results...)
- ⊞ [lang:cs] Romány a povídky (1 result...)
- ⊞ [lang:hr] Stranac (1 result...)
- ⊞ [lang:lt] Svetimas (1 result...)
- ⊞ [lang:es] El extranjero (46 results...)
- ⊞ [lang:ca] L'estrany (8 results...)
- ⊞ [lang:gl] O extranxeiro (1 result...)
- ⊞ [lang:de] Stranac (1 result...)

La Chute
Camus, Albert Claude (1913-1321)

- ⊞ [lang:en] The fall; and, The outsider (7 results...)
- ⊞ [lang:pt] A queda (11 results...)
- ⊞ [lang:fr] La chute (18 results...)
- ⊞ [lang:es] La Caída (8 results...)
- ⊞ [lang:ca] La caiguda (3 results...)

The Collected Fiction of Albert Camus
Camus, Albert Claude (1913-1321)

The Collected Fiction of Albert Camus (1960 - en)
Camus, Albert Claude (1913-1321); GILBERT, Stuart; O'BRIEN, Justin
[Printed Language Material] *Source: The British Library, ID: "000588158"*

(L'Étranger.) The Outsider
Camus, Albert Claude (1913-1321)

(L'Étranger.) The Outsider (1961 - fr)
Camus, Albert Claude (1913-1321); GILBERT, Stuart
[Printed Language Material] *Source: The British Library, ID: "000588183"*

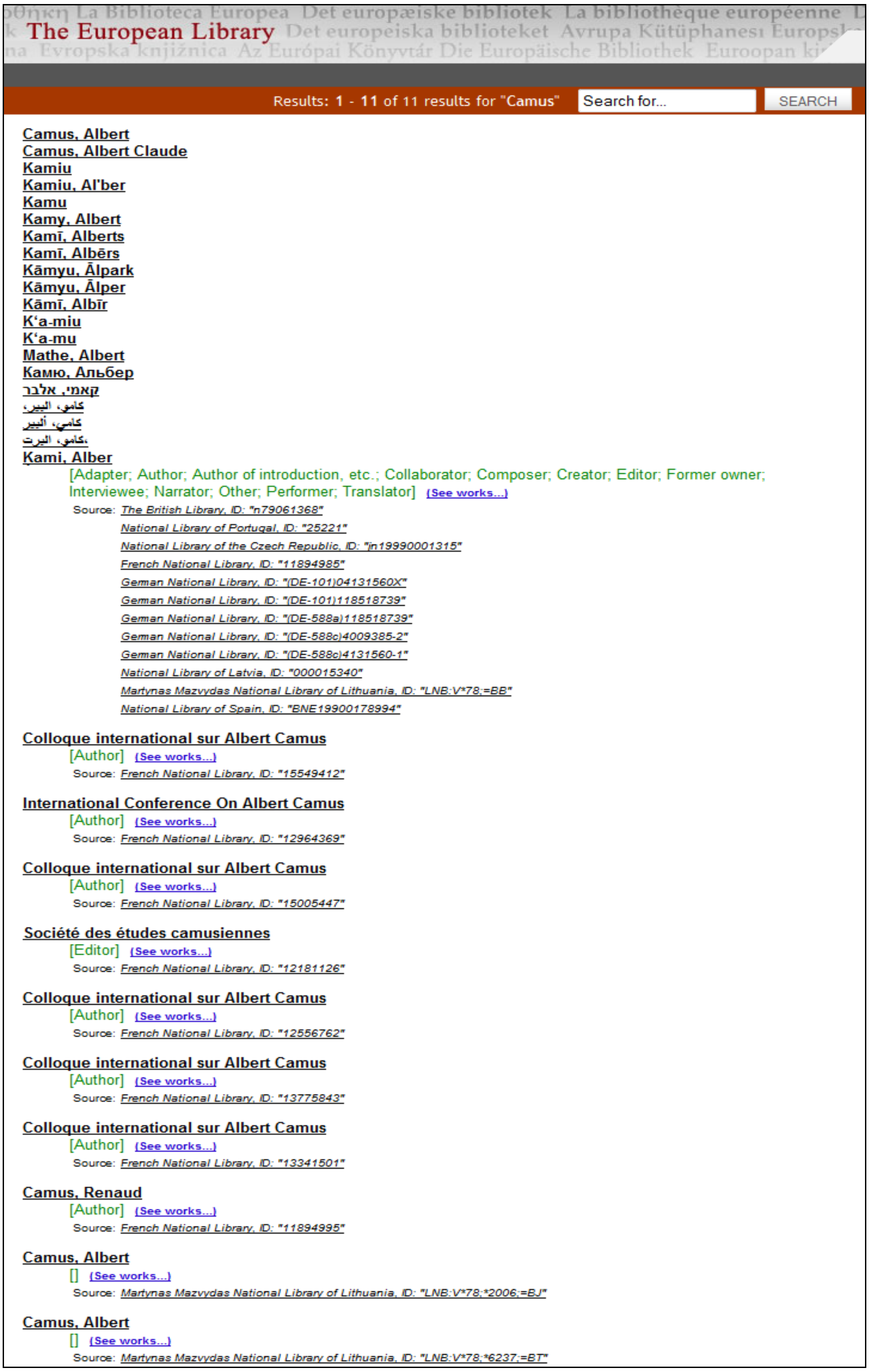
The Collected fiction
Camus, Albert Claude (1913-1321)

The Collected fiction (1960 - en)
Camus, Albert Claude (1913-1321); O'Brien, Justin; Gilbert, Stuart (1883-1969)
[Printed Language Material] *Source: French National Library, ID: "FRBNF329395320000004"*

The artist as outsider in the novels of Toni Morrison and Virginia Woolf
Williams, Lisa (1958-?)

The artist as outsider in the novels of Toni Morrison and Virginia Woolf (2000 - en)
Williams, Lisa (1958-?)
[Printed Language Material] *Source: French National Library, ID: "FRBNF388411680000000"*

Fig. 4. Search results for 'The Outsider' using the semantically enriched search interface.


βιβλιοθήκη La Biblioteca Europea Det europæiske bibliotek La bibliothèque européenne
The European Library Det europeiska biblioteket Avrupa Kütüphanesi Europsk
Evropska knjižnica Az Európai Könyvtár Die Europäische Bibliothek Euroopan ki
Search for...
SEARCH

Results: 1 - 11 of 11 results for "Camus"

Camus, Albert
Camus, Albert Claude
Kamiu
Kamiu, Al'ber
Kamu
Kamy, Albert
Kamī, Alberts
Kamī, Albērs
Kāmyu, Ālpark
Kāmyu, Ālper
Kāmī, Albīr
K‘a-miu
K‘a-mu
Mathe, Albert
Камю, Альбер
קאמי, אלבר
كامو، البير،
كامي، ألبير
كامو، البرت،
Ḳami, Alber
[Adapter; Author; Author of introduction, etc.; Collaborator; Composer; Creator; Editor; Former owner; Interviewee; Narrator; Other; Performer; Translator] (See works...)
Source: The British Library, ID: "n79061368"
National Library of Portugal, ID: "25221"
National Library of the Czech Republic, ID: "jn19990001315"
French National Library, ID: "11894985"
Geman National Library, ID: "(DE-101)04131560X"
Geman National Library, ID: "(DE-101)118518739"
Geman National Library, ID: "(DE-588a)118518739"
Geman National Library, ID: "(DE-588c)4009385-2"
Geman National Library, ID: "(DE-588c)4131560-1"
National Library of Latvia, ID: "000015340"
Martynas Mazvydas National Library of Lithuania, ID: "LNB:V*78;=BB"
National Library of Spain, ID: "BNE19900178994"

Colloque international sur Albert Camus
[Author] (See works...)
Source: French National Library, ID: "15549412"

International Conference On Albert Camus
[Author] (See works...)
Source: French National Library, ID: "12964369"

Colloque international sur Albert Camus
[Author] (See works...)
Source: French National Library, ID: "15005447"

Société des études camusiennes
[Editor] (See works...)
Source: French National Library, ID: "12181126"

Colloque international sur Albert Camus
[Author] (See works...)
Source: French National Library, ID: "12556762"

Colloque international sur Albert Camus
[Author] (See works...)
Source: French National Library, ID: "13775843"

Colloque international sur Albert Camus
[Author] (See works...)
Source: French National Library, ID: "13341501"

Camus, Renaud
[Author] (See works...)
Source: French National Library, ID: "11894995"

Camus, Albert
[] (See works...)
Source: Martynas Mazvydas National Library of Lithuania, ID: "LNB:V*78;*2006;=BJ"

Camus, Albert
[] (See works...)
Source: Martynas Mazvydas National Library of Lithuania, ID: "LNB:V*78;*6237;=BT"

Fig. 5. Search result for ‘Camus’ using the authority search interface.

5. Building a semantically richer model

The semantic cluster is loaded with the information resulting from the FRBRization of the bibliographic records. The bibliographic records are obtained from libraries voluntarily participating in TEL, as listed in Table 1. Both bibliographic and authority records from these libraries, encoded in UNIMARC or MARC21 formats, were processed for this purpose.

A first analysis of the source collections identified several data quality and normalization issues. For example, cataloguing practices followed by the libraries were quite heterogeneous. Reasons for this are usually related to the lack of proper support of cataloguing applications for the format in use (e.g. use of older versions), and also the inability of the format in use to keep pace with cataloguing needs. To deal with these heterogeneous cataloguing practices, an initial step was defined to **normalize** and assure the quality of bibliographic records for FRBRization. An overview of the challenges faced, along with an explanation of the chosen solution, is presented in [17].

The FRBRization of bibliographic data was thus performed in three steps (see Figure 6): normalization of the data, followed by entity extraction and entity aggregation.

The **entity extraction** step is responsible for extracting the semantic entities defined in the FRBR ontology from the bibliographic records. Several entities are generated from a single bibliographic record, since the entities defined in the FRBR ontology are defined in a more semantically rich model than the original MARC entities. For some of these entities only an abstract/simplified definition is extracted, given that the information provided in the bibliographic record is limited. This is the case for the concepts of Work and Expression, since bibliographic records are conceptually categorized as Manifestations and thus do not have a concrete correspondence. This step was done using a template containing rules for identifying elements (e.g. fields, subfields and data values) in the source record (encoded in MARC), transforming its data and creating new RDF entities defined in the FRBR ontology. The first version of the template was based on the FRBRizer tool [1],[18], which was extended for the purposes of this work.

Finally, in the **aggregation** step semantic entities are combined with entities identified in other bibliographic records until a complete graph is built. This is required because two bibliographic records may share the same entity with each other (e.g. both were created by the same author, share the same subject or were published by the same editor). The entities generated in the extraction step must be compared with each other to detect and merge duplicate references, until no duplicates exist in the ontology. To increase performance, the approach chosen was to develop a clustering algorithm that would group entities sharing a common set of characteristics and thus reduce the total number of comparisons required for duplicate detection. This clustering algorithm is applied first to Manifestations, then to Persons by looking at their relationship with Works (assuming that there are no two different authors of two different works sharing the same name and title), then to Expressions within the same work, and finally, to all Manifestations within the same Expression. Also the third group of entities (Concept, Object, Place and Event) is aggregated using the clustering algorithm.

Table 1

Processed collections from the Nobel Prize winners in Literature

Country Library		Bibliographic Records	Authority Records
UK	BL	18,838	3,020
Spain	BNE	12,444	2,294
France	BnF	28,302	286
Portugal	BNP	3,130	0
Germany	DNB	22,598	0
Belgium	KBR	1,081	0
Latvia	LNB	1,198	73
Czech Republic	NKP	2,945	231
Lithuania	NLL	341	56
Russia	NLR	20	7
Serbia	NLS	1,047	0
Hungary	OSKZ	2,887	3
	Total	94,831	5,970

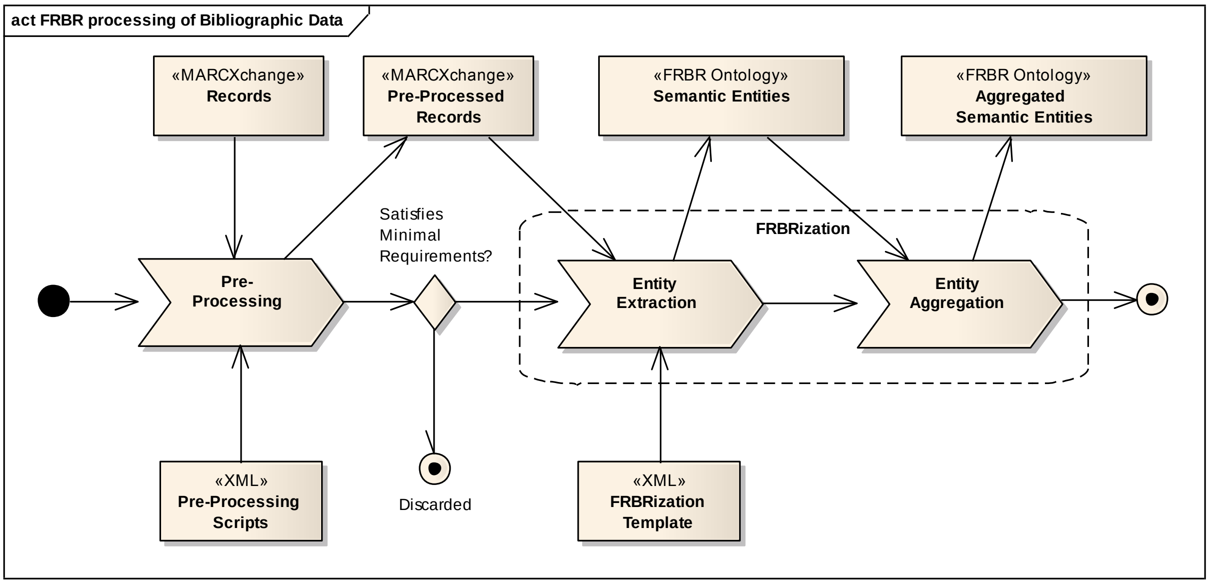

act FRBR processing of Bibliographic Data
«MARCXchange» Records
«MARCXchange» Pre-Processed Records
«FRBR Ontology» Semantic Entities
«FRBR Ontology» Aggregated Semantic Entities
Satisfies Minimal Requirements?
FRBRization
Pre-Processing
Entity Extraction
Entity Aggregation
«XML» Pre-Processing Scripts
Discarded
«XML» FRBRization Template

Fig. 6. Overview of the process used for the FRBRization of the records from the member libraries.

6. Evaluation of the results

Thirty-one volunteers performed a comparative evaluation of the two search services (the traditional OPAC and the semantically enhanced search interface). Invitations to evaluate the service were sent to the libraries that provided the original data, and to public mailing lists related to library users or professionals, along with instructions and an explanation of the purpose of the initiative. The respondents identified themselves as either library professionals or regular library users. The respondents were not given specific search tasks. Instead, they were first provided with a description of the test collections; then they were asked to perform any desired search tasks on both of the services; and finally they were asked to answer an online survey (this evaluation occurred over a week in December 2009).

The complete survey contained a total of ten questions: one question profiling the respondent; five questions evaluating the usefulness of FRBR for the semantic clustering of the results; and four open-ended questions requesting feedback for future directions in this work. The respondents were mostly library professionals, and their knowledge of OPAC, search engines and The European Library, are shown in Table 2. The results of the five questions that addressed the usefulness of the semantic clustering are summarized in Table 3. The responses received for the open-ended questions can be found in Appendix B. The respondents were asked to answer these questions while keeping in mind the comparison of their experience using the two services.

The general appreciation expressed by the respondents was positive, and the validity of the results is supported by their statistical analysis (P < 0,05 was obtained for questions 1, 2 and 4; P < 0,01 was obtained for questions 3 and 5). It is important to note that the question with the most positive feedback relates to the clustering of resources in different languages, stressing an interesting effect of the application of the FRBR model.

In general, the answers to the open-ended questions reinforce the positive feedback received from the survey, and most of the negative feedback can be related to problems with the original bibliographic data; minor software errors in the prototype; or functionalities that would be expected in a final service but that were intentionally not implemented, given that it was only a prototype.

Additional points can be highlighted in the analysis of the answers. Many users seem to have performed search tasks centred on a particular author. Some answers indicate that the semantic clustering was helpful for these author-centred tasks, and some suggest further improvements to support them (for example, to identify works by and about an author). Further semantic relations and data were requested for improving the quality of the semantic clusters and their navigation (mainly regarding expressions in different languages and media types). Although the original bibliographic data limits what can be done, these answers reinforce the usefulness of semantically richer organization of search results and point to further areas in which the service might be developed.

The analysis of the survey supports the hypothesis that FRBR clustering is feasible with existing data, opening doors for its application to any existing traditional bibliographic system.

Table 2

Results for the respondent profiling question (multiple choice)

Question: Which of these statements apply to you?	**Response count**
I'm a library professional	21
I'm familiar with searching in libraries' online catalogues	15
I'm familiar with searching in The European Library	3
I'm familiar with searching in Internet search engines	17

Table 3

Results for question addressing the usefulness of the semantic clustering

Question	**Yes, always**	**Yes, sometimes**	**No**
Is the clustering of results helpful for finding the relevant resources for your queries?	15	12	3
Is the clustering of results helpful for discovering additional relevant resources to what you were looking for?	11	16	3
Is the clustering of results helpful for discovery of resources in different languages?	21	8	2
Is the clustering of results helpful for finding the first publication of a resource?	5	17	9
Is the clustering of results helpful to discover which libraries hold the same resource?	13	15	3

7. Conclusion

The work described tested the integration of new features provided by web search engines – single search, relevance ranking, and relevance feedback - into a traditional OPAC by taking advantage of a semantically richer organization of the data available in common bibliographic records. Evaluations performed by library users and library professionals using the semantically enriched interface were clearly positive, in particular praising its ability to support the clustering of related resources in different languages.

Most of the negative feedback was related to minor software errors of the prototype, or functionalities that would be expected in a final service but that were consciously not implemented in the prototype. Another important reason for negative feedback is related to problems with the original MARC data that could not be solved by the approaches discussed in this paper. We think that these problems are the biggest challenges when migrating from a traditional bibliographic catalogue to an FRBR-based catalogue, and will require a significant effort from librarians and researchers.

The results of this experiment identified new research opportunities. The **relevance ranking** of manifestations within the same cluster (work) could be further explored, since the user may not be looking for the first publication of a work, but instead, for a particular expression in a specific language. For example, correlations with other properties of the manifestation, such as the publication date and language, could be also explored. **Clustering** of superworks (a work being a bibliographic antecedent of several other works), aggregated works (sets) and serial works were not given detailed attention in this study. The clustering of these types of works is a big challenge, both algorithmically and visually, but has the ability to greatly improve the usability of an OPAC.

This experiment supports the hypothesis that semantically richer models can be built from existing bibliographic data, and can be effectively used to improve the user experience when searching and browsing bibliographic catalogues.

Acknowledgments

This work was supported by FCT (INESC-ID multiannual funding) through the national PIDDAC Program funds, and by the European Commission through the TELplus project (contract ECP-2006-DILI-510003).

References

[1] Allgood, J. E. (2007). Serials and Multiple Versions, or the Inexorable Trend Toward Work-Level Displays. Library Resources & Technical Services, 51(3), 160-178.

[2] Young, M., & Yu, H. (2004). The Impact of Web Search Engines on Subject Searching in OPAC. Information Technology and Libraries, 23(4), 168-180.

[3] IFLA Study Group on the Functional Requirements for Bibliographic Records. (1988). Functional requirements for bibliographic records: final report. UBCIM publications, new series, 19. München: K. G. Saur. Available: http://www.ifla.org/VII/s13/frbr/index.htm.

[4] Hegna, K., & Murtomaa, E. (2003). Data Mining MARC to find FRBR? International Cataloguing and Bibliographic Control, 32, 52-55.

[5] Ayres, M. L., Kilner, K., Fitch, K., & Scarvell, A. (2002). Report on the Successful AustLit: Australian Literature Gateway Implementation of the FRBR and INDECS Event Models, and Implications for Other FRBR Implementations. Available: http://www.ifla.org/IV/ifla68/papers/054-133e.pdf.

[6] Yee, M. M. (2005). Frbrization: a Method for Turning Online Public Finding List into Online Public Catalogue. Information Technology and Librarians, 24(3), 77-95. Available: http:// repositories.cdlib.org/postprints/715.

[7] Salaba, A., & Zhang, Y. (2007). From a Conceptual Model to Application and System Development. Assis&t – The Information Society for the Information Age, Bulletin, August/September 2007. Available: http://www.asis.org/Bulletin/Aug-07/salaba_zhang.html.

[8] Žumer, M. (2007). FRBR: The End of the Road or a New Beginning. Assis&t – The Information Society for the Information Age, Bulletin, August/September 2007. Available: http://www.asis.org/Bulletin/Aug-07/Zumer.pdf.

[9] Aalberg, T. (2006). A process and tool for the conversion of MARC records to a normalized FRBR implementation. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (ed.) ICADL 2006. Lecture Notes in Computer Science, 4312, 283–292. Springer.

[10] Pisanski, J., & Žumer, M. (2007). Functional Requirements For Bibliographic Records: An Investigation Of Two Prototypes. Electronic Library & Information Systems, 41(4), 400-417.

[11] Dellit, A., & Boston, T. (2007). Relevance ranking of results from MARC-based catalogues. National Library of Australia Staff Papers. Available: http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1052.

[12] Khoo, C., Poo, D., Teck-Kang, T., & Hong, G. (1999). E-Referencer: Transforming Boolean OPACs to Web Search Engines. IFLA Council and General Conference, 65. Available: http://archive.ifla.org/IV/ifla65/papers/010-143e.htm.

[13] Antelman, K., Lynema, E., & Pace, A.K. (2006). Toward a 21st Century Library Catalog. Information Technology and Libraries, American Library Association, 128-139.

[14] Michalko, James, & Haeger, John. (1994). The research libraries group: Making a difference. Library Hi Tech, 12(2), 7 - 32

[15] Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5, 199–220.

[16] Bekiari, C., Doerr, M., & LeBoeuf, P. (2009). FRBR: Object-Oriented Definition and Mapping to the FRBR-ER. Version 1.0. International Working Group on FRBR and CIDOC CRM Harmonisation. Available: http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBRoo_V1.0_2009_june_.pdf.

[17] Manguinhas, H., Freire, N., Borbinha, J. (2010). FRBRization of MARC records in multiple catalogs. JCDL 2010, 225-234.

[18] Aalberg, T., & Žumer, M. (2008). Looking for Entities in Bibliographic Records. ICADL (2008). Lecture Notes in Computer Science, 5362, 327-330. Springer, Heidelberg.

Appendix A – Overview of the FRBR entities and attributes found in the bibliographic records after the entity extraction step of FRBRization

Class / Property	total	with	max	avg
Work	92444	-	-	-
antecedent	2123	1736	15	0.02
antecedentOf	2123	2123	1	0.02
creationDate	10	10	1	0.00
creator	89577	87897	15	0.97
form	33691	27905	6	0.36
intendedAudience	29328	28893	3	0.32
language	61088	60988	4	0.66
notCreator	3	3	1	0.00
partOf	15465	15465	4	0.17
realization	94117	90067	36	1.02
subject	26041	21249	8	0.28
subjectOf	252	252	1	0.00
title	10288	10282	2	0.11
variantTitle	11	11	1	0.00
Expression	94117	-	-	-
contentSummarization	807	657	24	0.01
edition	2998	2998	1	0.03
editionStatement	20616	20484	2	0.22
embodiment	90117	90092	2	0.96
extent	133	133	1	0.00
language	79763	78807	7	0.85
performanceMedium	213	127	7	0.00
publicationDate	31127	30936	5	0.33
realizationOf	94117	94117	1	1.00
realizer	55393	35961	109	0.59
relatedEdition	39	27	4	0.00
scale	3	3	1	0.00
title	35542	35542	1	0.38
translation	4011	4011	1	0.04
translationOf	4011	3799	35	0.04
typeScore	1932	1271	2	0.02
Manifestation	90796	-	-	-
accessAddress	1	1	1	0.00
alternativeID	49704	39861	40	0.55
artifact	90661	90661	1	1.00
captureMode	564	564	1	0.01
carrierAccompanyingMaterial	364	355	3	0.00
carrierCartographic	5	5	1	0.00
carrierDimensions	21204	20954	4	0.23
carrierElectronic	148	148	1	0.00
carrierExtent	27683	27550	3	0.30
carrierFilm	34	32	2	0.00
carrierGraphics	224	224	1	0.00
carrierMedium	29929	29529	3	0.33
carrierMicrofilm	2	2	1	0.00
carrierSound	1067	1067	1	0.01
cuttingKind	25	25	1	0.00
distributor	597	593	2	0.01
editionStatement	23712	23484	2	0.26
editionType	62	62	1	0.00
embodimentOf	90117	90092	2	0.99
fabricationDate	413	413	1	0.00
fabricationPlace	7251	7227	4	0.08
grooveWidth	191	191	1	0.00
imageColour	694	666	2	0.01
otherTitle	2162	1611	20	0.02
playingSpeed	398	398	1	0.00
presentationFormat	432	250	2	0.00
producer	912	610	8	0.01
publicationDate	88769	88769	1	0.98
publicationPlace	214400	90647	19	2.36
publisher	6162	4847	10	0.07
reproductionCharacteristics	262	262	1	0.00
responsibility	85153	69204	45	0.94
soundKind	25	25	1	0.00
subtitle	35112	34108	8	0.39
systemRequirements	7	7	1	0.00
tapeConfiguration	111	111	1	0.00

Class / Property	total	with	max	avg
title	90645	90645	1	1.00
variantTitle	7773	6107	41	0.09
Person	167626	-	-	-
creatorOf	89288	89144	2	0.53
date	129660	129643	2	0.77
name	167619	167617	2	1.00
notCreatorOf	3	3	1	0.00
other	2135	2082	2	0.01
producerOf	258	258	1	0.00
publisherOf	5163	5163	1	0.03
realizerOf	53866	53751	2	0.32
role	97759	96345	6	0.58
subjectOf	4312	4312	1	0.03
surname	71064	71062	2	0.42
Corporate Body	3829	-	-	-
creatorOf	289	289	1	0.08
distributorOf	597	597	1	0.16
name	3829	3829	1	1.00
producerOf	654	654	1	0.17
publisherOf	998	998	1	0.26
realizerOf	1527	1527	1	0.40
role	2928	2928	1	0.76
subjectOf	250	250	1	0.07
Concept	21268	-	-	-
scheme	21268	21268	1	1.00
term	17802	17802	1	0.84
Place	279	-	-	-
city	144	144	1	0.52
country	152	152	1	0.54
date	202	163	3	0.72
venue	141	141	1	0.51

Legend: Total number of classes/properties (total); Number of classes with at least on occurrence of a specific property (with); Maximum number of occurrences of a given property found for each class (max); Average number of occurrences of a given property found for each class (avg).

Appendix B – Results of the open questions of the evaluation survey

Question: Can you describe other user tasks for which the clustering of results is helpful?

The following comments were collected:

1. Clustering could also be helpful to identify candidates for duplicates.
2. Which authors have a work in the list?
3. Many clustering would be helpful. So the possibility for the end-user to choose different clustering method would be useful. E.g. the OPAC would allow the user to choose clustering by language, by country, by date, etc.
4. Finding all the books by same author
5. Get the whole bibliography of an author
6. The clustering of results helps organizing the search and offers the possibility to find content of different type (written, audio etc.)
7. The result to get an overview of an author's complete production
8. Showing many results in a single page, occupying less space on the displayed page (limited scroll)
9. I don't really know. It would be easier to browse the clusters if there were some reference by which criteria the results are grouped.
10. It makes the resource identification faster and more precise. Minimizes the noise in the search result display, and avoid the need of skimming through a large amount of records.
11. Actually the clustering can help to find resources of a particular media type (audio book, print, Braille etc.) but unfortunately your user interface and your search form doesn't provide this functionality.
12. getting more relevant results which is helpful in database contents
13. My problem with the OCLC FictionFinder is that it looks too much to aim at ‘literate undergraduate’: I would expect this service to meet requirements also of ‘literate postgraduates’ - researchers who need more in-depth information like, as your question states, 1st edition, translation of the 3rd revised edition in a particular language, and the like; also notes: keep the notes in the records!!!
14. for finding the same work in different types of materials
15. The clustering of results should enable the user to select a specific resource. It is not clear that "printed language material" would be adequate

for this task. The user will want to know whether the resource is a book, an e-book, a sound recording, a Braille text, etc.

Question: What did you LIKE the most in the clustering of results?

The following comments were collected:

1. The languages are clearly identified. The gaps in the clustering show clearly how much still has to be done in cataloging, more 130s / 240s, and more roles in 1XX and 7XX $e + $4, and 250s with more standardized strings. I was glad to see non-roman scripts, at least Cyrillic was easy to be found. Very impressive! The MARC record in xml makes it very easy to analyze the structures behind the clusters. Not a feature for the end user, but for the expert.
2. Fewer hits.
3. Seeing the resources of a particular author by language
4. No so many pages found with the same title
5. clearly represented results page, often short list of results
6. I liked that it made simpler the process of searching for information: - Quick translation of each title into different languages - Displayed number of results by different languages - The icon explaining the printed text or multimedia
7. Being able to go view an author’s work knowing that all results related to that particular author are showed on that section of the page.
8. I wanted to have option to select criteria by which grouping the results. I wanted to have opportunity to see the result list in various different views.
9. It allows multilingual search, groups related resources, and different formats.
10. The presentation of search results
11. that I overviewed all the search results this quick and the possibility to explore instead of scroll
12. relevant results
13. This is definitely a very good start: the difference between the clustered and unclustered records shows the value of the service.
14. the "condensing" the presented information, which allows one to spot easily the search results that are really distinct (and not to bother right away with dozens of translations and new editions in the results)
15. nothing
16. Is the first application I've seen which really tries to implement FRBR.
17. The number of results of the query is more limited and easier to browse
18. I found transparent, correct and useful metadata in different catalogues with one click.
19. The basic principle of returning a structured result set is powerful and much better than a simple list of all results by date, for example.
20. The opportunity to see all the editions (in the same language and all the others etc.) for every work. It is very helpful.
21. The fact that it aggregates records in multiple languages...

Question: What did you DISLIKE the most in the clustering of results?

The following comments were collected:

1. There seems to be no authority control, e.g. for the creators: Searching for "Mistral" brings up an unsorted list of works by "Mistral, Frédéric (1830-1914)" or "Mistral, Gabriela (1889-1957)", mixed into one list. So it is visible that before extracting the FRBR group 1 entities, there would have to be the group 2 entities, controlled, in order to make the clustering more efficient. Non-sorting parts of a title seem to break the clustering, e.g. "Die Blechtrommel" doesn't seem to match with "-Die -Blechtrommel".
2. The simple search gave (expectedly) much poorer results, e.g. doing a search for Toni Morrison in the simple search instead of in the author field in the advanced search
3. some wrong titles where in the cluster
4. Not all manifestations of a work are clustered probably because of different cataloging rules or practice. The clustering seems to be mainly based on the same strings of uniform titles, title fields and subfields. Probably it would be helpful to adjust the algorithm so that criteria such as same author and translator and mainly the same title: e.g. the following titles didn't match because the remainder of the first title and the statement of responsibility are not labeled with a subfield code.
5. The inability to handle diacritics. XML record on click on the title (not user friendly)
6. That the first record of the result set wasn't what exactly what I searched for. If I am searching for concrete author or title I am expecting to retrieve items.

7. If an author / title search is launched (e.g. "Platero y yo (Title)" and "Juan Ramon Jimenez (Author)") the systems shows not just this works, but also others non related items (maybe because the default operator is OR?) MARCXML visualization is not easy.
8. time consuming
9. Not "dislike" but: there is a mixture of presented records/clusters ordered - to the user, in an unidentified way.
10. The quality of results set is sometimes bad. But maybe that was rather due to query disambiguation issues. Other times the clustering is much incomplete, however (translations that were not aggregated)
11. everything
12. Unclustered items, i.e., expressions of work not grouped with the corresponding work. Editions in different languages are unsorted, or I couldn't find the criteria. Original language versions must appear first. Actually, there's no indication about which is the original language. The whole result set must be sorted by original language too, that way real work records would be at the top while errors in clustering would appear last. The date of the work is not displayed. Instead, the display gives primacy to the year of publication, which is not so important when looking for works. I think the difference between expressions and manifestations are not considered. It's not very easy to find this difference with the selection used on the prototype.
13. Not all the works are correctly clustered,
14. They didn't really seem to cluster well. English and French editions of Peal S. Buck's The Mother were separate entries.
15. The FRBRisation process only seems to work for titles. It does not work for authors.
16. I do not see the whole record, all the metadata for a book...
17. The interface stills a bit confusing...

Question: Do you have additional suggestions, comments or proposals?

The following comments were collected:

1. Is there an order in the list of search results? I would very much like to read the other colleagues opinions.
2. This is cool! I really like it! I would be nice to have an ajax like auto-complete from the simple search box, e.g. if I type Toni M... it fills the rest in for me - this would help improve the simple search. It would be nice to see the full bibliographic records. It would be nice to link to the record in the local library interface. It would be nice to link to the full text!!! It would be nice to see a logo for the local library? It would be nice to have some help texts to help a user use the service. It would be nice to have a second cluster by date, e.g. Sula / Toni Morrison (currently listed by language and then within language you get the dates) however, if you could also see all the Sula's published in for example 1985
3. nice to see that the clustering of titles with good metadata (such titles with uniform titles even in different fields e.g. Unimarc: 454, 304, Marc: 240, 730, 594) works really good
4. I think clustering of results is an interesting approach for searching through The European Library
5. Works by and about an author should be presented in the one search but separated in the search result
6. It would be useful to have opportunity to select records by subject list. More useful if I could select records by subject list in my native language.
7. An FRBR-layered search visualization might begin firstly with the identification of the author, instead of directly displaying the works, for validation purposes. Once the user identifies the author properly, then his/her works would be shown. The same goes for the rest of search options (title, subject...)
8. I would like to have little icons next to expandable results on language level which filter media types or depict the existence of various media types
9. to increase the data base more relevantly according to searches
10. I would suggest the service to make distinction between the works by the author and works about the author: in the present display you have to scroll/browse to find the works by the author because you somehow cannot believe that the works you get under the author's name are the only ones held in libraries. Obviously you will have to work some more on the authority data. My search was Pablo Neruda. Let Serbian NL check its record on Pablo Neruda (lang: hr, but it is really Slovenian!!)
11. FRBR is incomprehensible – give it up
12. FRBR shows it power when dealing with other carriers than just books. Perhaps you could try

with movies related with novels in your prototype.

13. Clustering of results for authors as well as titles would be very desirable... For example, "Kipling" returns an undifferentiated set of results, which is a mixture of works by Kipling, critical works and translations. An FRBRised display should distinguish between works by the author; expressions (e.g. translations) of works by the author and works about the author. It would also be desirable to identify derivative works, but existing data may make this difficult.
14. To make available the whole record, including holdings
15. I really dislike the fact that I cannot search random. Why do I need to select one country? I look for an object but I don't know the location of it, or I want to see all data of this object.