The Ontology-based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services

E. Francesconi[1], M.W. Küster[1], P. Gratz[2], and S. Thelen[2]

[1] Publications Office of the EU
[2] infeurope S.A.

Abstract. The Publications Office of the European Union is responsible to make available and disseminate the official publications and bibliographic resources produced by the institutions of the European Union. The central component of its information system is the CELLAR repository, providing semantic indexing, advanced search and data retrieval for multilingual resources. This paper gives an overview of the semantic modeling approach for CELLAR, based on semantic web technologies. Moreover, a proposal for a possible evolution aiming to improve the modularity and facilitating the general management of the model is shown.

Keywords: Multilingual documents, Semantic indexing, Knowledge modeling.

1 Introduction

The dissemination of the official documents as well as other bibliographic resources produced by the European Union institutions is the main mandate of the Publications Office of the European Union, in its role of inter-institutional body of the European Commission in charge to inform about government procedures and legal framework.

As for official publications[3], this right is guaranteed by the availability of such resources in the 24 official languages spoken in the 28 member states of the European Union, while other publications (like tendering documents, general publications and information on EU-funded research projects) are mainly available in the 3 fundamental languages (English, French and German). With the authentic and legally binding publication of the electronic edition of the Official Journal (e-OJ) from 1 July 2013, the on-line accessibility of legal resources has become an essential requirement, guaranteed by the Eur-lex service[4].

[3] Treaties, International agreements, Legislation, Complementary legislation, Preparatory acts, Case-law, National implementing measures, References to national case-law concerning EU law, Parliamentary questions, Consolidated legislation, Other documents published in the Official Journal C series, EFTA documents

[4] http://eur-lex.europa.eu

In the last couple of years most efforts of the Publications Office (OP) were focusing on a project aiming to transform the archival and dissemination architecture, based on different systems, into a federative architecture based on a common archival service, providing also a common interface for disseminating materials to the users (Fig. 1). The central component of this architecture

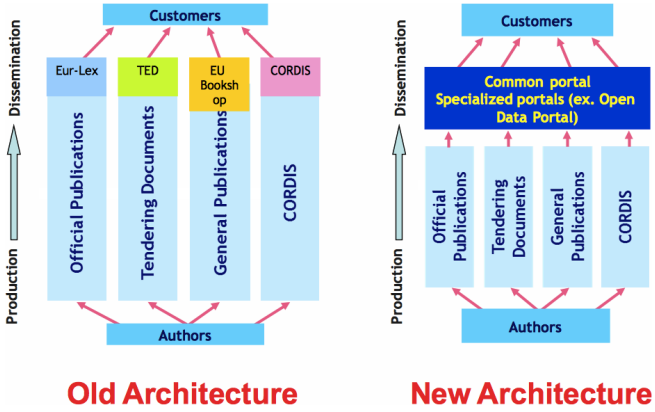

Fig. 1: The Publications Office archival and dissemination transformation programme

is CELLAR, a content and metadata repository containing documents coming from the production and postproduction services (including content validation and metadata production). They are available for long term preservation, open data, indexing, as well as advanced search and retrieval services. CELLAR resources are semantically described by an ontology, which represents the Common Metadata Model (CDM) of the OP resources. This paper is focused on the description of this ontology as well as on possible developments. It is organized as follows: in Section 2 the architecture of the CELLAR platform is described; in Section 3 the CELLAR multilingual semantic approach, represented by CDM, is presented; in Section 4 a possible evolution of CDM is illustrated and in Section 5 the advantages of the proposed evolution are discussed. Finally in Section 6 some conclusions are reported.

2 The CELLAR architecture

CELLAR represents the central hub of the whole information system of the OP. It is based on a Fedora digital objects repository[5], organized in two logical units including Oracle database technologies: content is stored in the CELLAR

[5] http://www.fedora-commons.org

Common Content Repository (CCR) currently[6] including about 152 million documents in 24 languages; metadata in as many languages are stored in the CELLAR Common Metadata Repository (CMR) described by semantic web technologies, resulting in about 1100 million triples, stored in an RDF triple store. Currently CELLAR receives about 5 million requests per day, providing information results for the EUR-Lex service and for the query service (SPARQL endpoint) recently exposed in order to complement linked open data services to potential consumers. Other services and types of resources, like TED for tendering documents, EU Bookshop for general publications, CORDIS for information on EU-funded research projects will be served by CELLAR in the near future (Fig. 2). Concerning disaster recovery and emergency management a proper data replication service for the production database has been put in place as shown in Fig. 2.

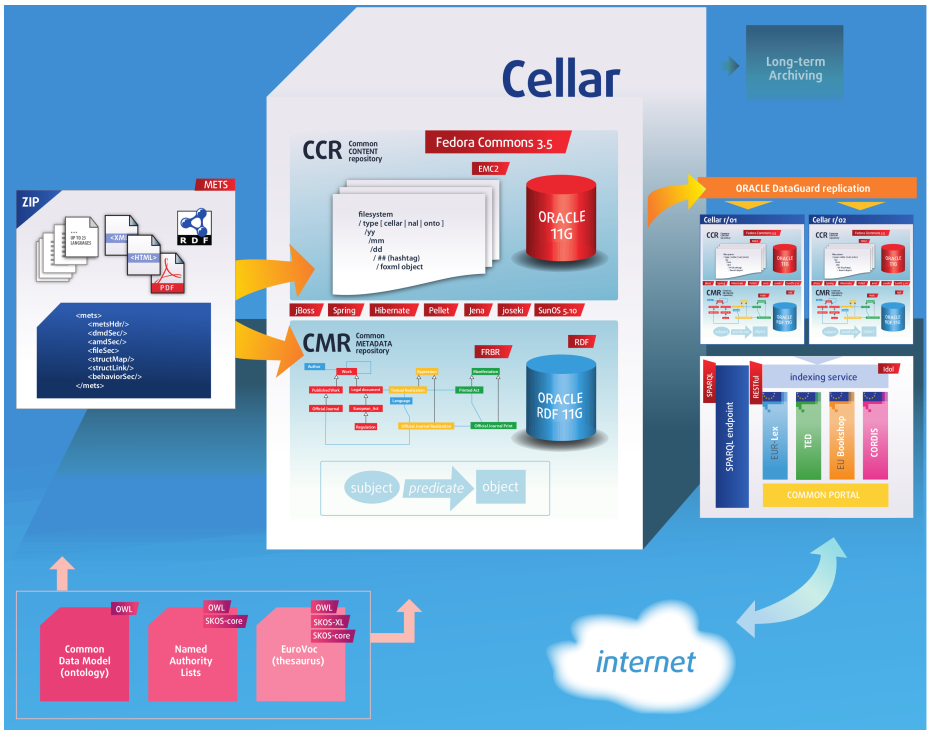

Fig. 2: The CELLAR architecture and services[7]

Based on CDM, the Common Metadata Repository (CMR) represents the essential asset to guarantee multilingual semantic access services to the CELLAR

[6] March 2015
[7] courtesy F. Sanmartin

contents. The following section depicts how the CDM allows to describe, from a semantic point of view, all the OP resources.

3 Common Metadata Model (CDM)

The current CDM is an ontology based on the FRBR[8] model [1], described by RDF(S)/OWL technologies, able to represent the relationships between the resource types managed by the OP and their views according to the FRBR model in terms of Work, Expression, Manifestation and Item. In the current CDM organization, the FRBR hierarchy represents a sort of pivot knowledge organization system, according to which resource types (general publications, legal resources, legislation, case law, etc.) and FRBR views (ex: general publication expression, case law expression, official journal manifestation, etc.) are organized through sub-class relationships (Fig. 3).

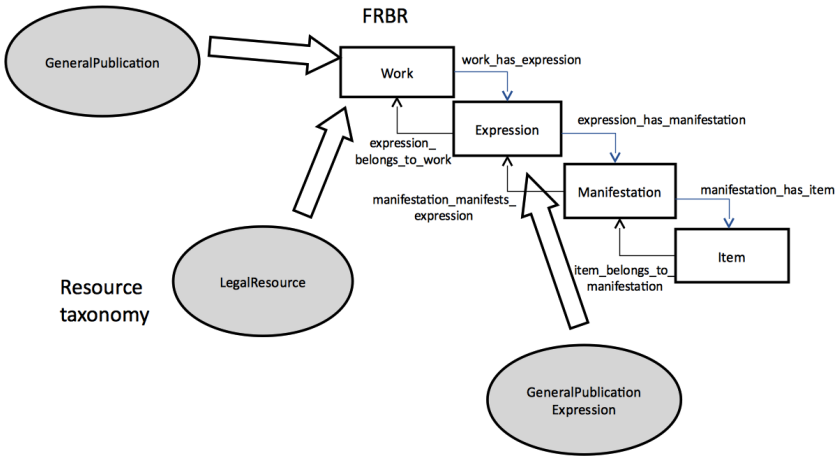

Fig. 3: The current CDM organization system

Resources are identified by URIs classified according to the FRBR hierarchy, thus organizing the objects managed by CELLAR at different FRBR abstraction levels. Such URIs have for example `http://publications.europa.eu/resource/oj/` as namespace of the official journal resources, followed by an ID created as concatenation of metadata values at each FRBR level (see Tab. 1 for some examples).

Based on commonly known best practices for linked data, CELLAR enables clients to retrieve various resource representations via content negotiation [5]. Each FRBR entity is represented by a non-information resource [6, 7] that redirects client requests to one of its different representations depending on the

[8] Functional Requirements for Bibliographic Records

Resource FRBR type	Resource ID
Work (Regular OJ publication n. 26 of 2015)	JOL_2015_026_R_0001
Expression (english variation)	JOL_2015_026_R_0001.ENG
Manifestation (PDF/A-1a format)	JOL_2015_026_R_0001.FRA.pdfa1a
Item	JOL_2015_026_R_0001.FRA.pdfa1a.l_02620150131fr00010002.pdf

Table 1: URIs at different FRBR levels

HTTP headers (i.e. Accept, Accept-Language, etc.) specified by client. Thus, in order to retrieve an RDF/XML representation of Work-level metadata a client has to perform an HTTP get request against the non-information resource of the Work like follows:

```
curl -L -H "Accept:application/rdf+xml"
http://publications.europa.eu/resource/oj/JOL_2015_026 R_0001
```

Equivalent requests can be triggered against Expression or Manifestation URIs to yield metadata from these FRBR levels. Moreover, Cellar provides also different aggregated representations like tree or branch notices that contain metadata from all different FRBR levels. These aggregated representations are also retrievable by issuing HTTP requests against the non-information resources representing a Work/Expression aspect of the bibliographic entity. Consequently, those resources are not only an access point for bibliographic information that belongs to a specific FRBR aspect, but also a proxy for getting a representation of the complete bibliographic entity. The approach proposed in the following section allows to tackle this asymmetry.

Of all existing representations, the so called tree notice of a FRBR hierarchy is the representation that best describes the complete bibliographic record from a "web of data" point of view, since it provides the entire set of metadata at each level of the FRBR hierarchy in a single RDF serialization. This representation can be retrieved via a HTTP get request against the Work like follows:

```
curl -L -H "Accept:application/rdf+xml;notice=tree"
http://publications.europa.eu/resource/oj/JOL_2015_026 R_0001
```

This CDM version is currently in production providing detailed views, in particular regarding language versions and formats, of the OP resources, for both documents and metadata search and retrieval services, as well as for the OP common portal.

In the context of a recent activity a review of the current CDM in order to reduce complexity of the query framework was performed. During this review the following shortcomings of the current model have been revealed:

1. the mixture in the same taxonomy of resource types and FRBR classes
2. the need to follow complex paths to reach different FRBR views of the same resource type (see General Publication type: GeneralPublication → Work → Expression → GeneralPublicationExpression).

These issues result in certain limitations of the framework. For instance that, given a resource type, the access to the different levels of the FRBR model is not direct. Moreover, it is necessary to know the type of a resource at query level in order to retrieve metadata at each level of the FRBR model, while it would be more simple that, given a resource, there is a common query to access metadata at different FRBR levels, irrespective of the resource type.

In the next sections an overview of the current discussion about a possible CDM evolvement is presented.

4 Proposal for CDM evolvement

A proposal for possible evolvement of the current CDM approach aims firstly to keep a distinction between the taxonomy of the resources and the FRBR model.

A *Resource* in the ISBD[9] sense is defined as "*an entity, tangible or intangible, that comprises intellectual and/or artistic content and is conceived, produced and/or issued as a unit, forming the basis of a single bibliographic description*". Therefore, resources are actually not equivalent to, or sub-class of, any individual FRBR classes [2]. As pointed out in [3] each FRBR classes *reflects* one aspect of a resource, seen as a bibliographic entity at different levels of abstraction.

A *Resource* (in the ISBD sense) has the same intention as the combined attributes of the FRBR model [3], therefore it can be considered as the disjoint union of the Work, Expression, Manifestation and Item levels in FRBR model, as expressed by (1).

$$Resource = Work + Expression + Manifestation + Item \tag{1}$$

The relationship between the two domains (resource taxonomy and FRBR model) is therefore of *part-of / aspect*. In this context, every FRBR level is an *aspect* of a current resource and can be considered as collector of the metadata able to describe a resource at that level.

Therefore, a resource and its FRBR model can be viewed as aspects of the same reality in two perspectives [2]:

1. The "web of data" perspective
2. The "bibliographic data" perspective

A resource identified by a specific URI represents an entity of the "web of data". The resources published by the OP are basically bibliographic entities. Therefore, they can be described according to the FRBR model. Works, Expressions, Manifestations and Items of the FRBR model are also type of entities of the web of data, but they can also be viewed as a specific aspects of a bibliographic resource, therefore viewed in the "bibliographic data" perspective.

This distinction provides the main motivation for improving CDM with the goal to simplify the query framework, thus improving the accessibility of the resources. To achieve this goal, the following actions have been undertaken:

[9] International Standard Bibliographic Description

1. Introduction of a logical separation between the taxonomy of the OP resources and the FRBR model, therefore avoiding any subClass relations between them;
2. Introduction of cdm:has[FrbrClass]Aspect[10] relations between a classes of the OP resource taxonomy and their aspects as FRBR classes (e.g.: cdm:hasWorkAspect, cdm:hasExpressionAspect, etc.);
3. Introduction of a rdfs:subPropertyOf relation between cdm:has[ResourceTypeFrbrClass]Aspect at different levels of the taxonomies.

In Fig. 4 a sketch of the OP resource taxonomy (limited, for simplicity, to the root and one subclass) and its relationships with the FRBR model at each taxonomy level is represented. In particular the generic class of OPBibliographicResource is linked with cdm:hasWorkAspect, cdm:hasExpressionAspect, cdm:hasManifestationAspect, cdm:hasItemAspect to the corresponding classes of the FRBR model. Sub classes in the resource taxonomy, like SourceOfLaw, are linked to the corresponding classes of the FRBR model with similar specific properties (as cdm:hasSourceOfLawWorkAspect, cdm:hasSourceOfLawExpressionAspect, etc.). Such "aspect" properties are organized in pure taxonomic relationships (subPropertyOf) for each level of the FRBR model (cdm:hasSourceOfLawWorkAspect is a sub property of cdm:hasWorkAspect, and so on).

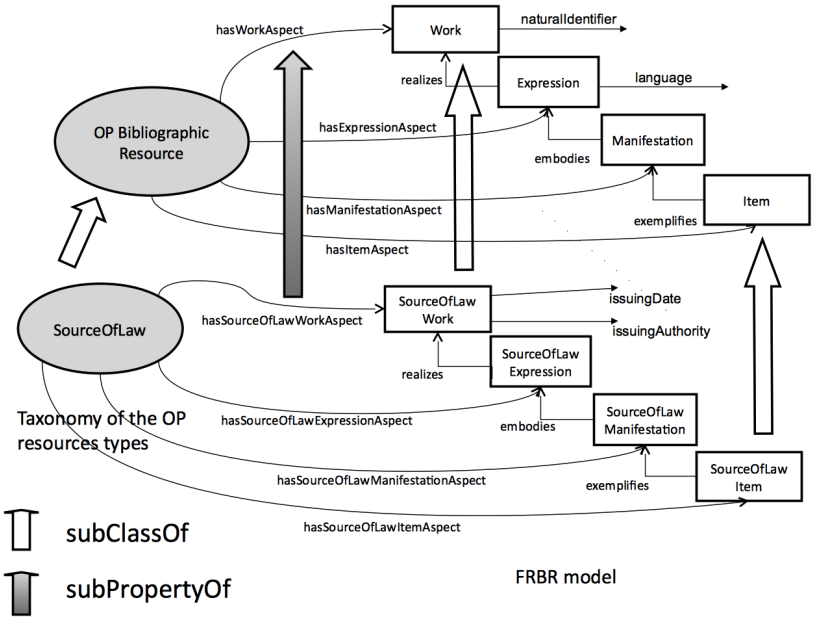

Fig. 4: A proposal for CDM organization system development

[10] where cdm: is the CDM namespace and [FrbrClass] is one of Work, Expression, Manifestation or Item classes

The FRBR classes are collectors of resource metadata at their specific taxonomy level: for example (see Fig. 4) at Work aspect level, a resource will have for example cdm:naturalIdentifier as generic metadata, described by object or datatype properties, shared by all the OP bibliographic resources. Similarly at SourceOfLawWork aspect level, specific metadata shared by all the sources of law, are given, as for example cdm:issuingDate and cdm:issuingAuthority of a legal measure. The same holds for the other FRBR classes at each level of the OP resource taxonomy.

In this CDM model, the resource identifiers representing the complete record are derived from the cellar tree notice URI `cellar:[hash-value]/rdf/tree/full` [11] as follows `cellar-record:[hash-value]` [12]. The following RDF snippet depicts the usage of this URI.

```
<rdf:Description rdf:about=
 "cellar-record:58da3a99-a91d-11e4-8e01-01aa75ed71a1">
 <rdf:type rdf:resource="cdm:SourceOfLaw"/>
 <cdm:hasSourceOfLawWorkAspect
     rdf:resource="ojns:JOL_2015_026_R_0001"/>
 <cdm:hasSourceOfLawExpressionAspect
     rdf:resource="ojns:JOL_2015_026_R_0001.ENG"/>
 <cdm:hasSourceOfLawExpressionAspect
     rdf:resource="ojns:JOL_2015_026_R_0001.FRA"/>
 <cdm:hasSourceOfLawManifestationAspect
     rdf:resource="ojns:JOL_2015_026_R.ENG.pdfa1a"/>
 <cdm:hasSourceOfLawManifestationAspect
     rdf:resource="ojns:JOL_2015_026_R.FRA.pdfa1a"/>
 <cdm:hasSourceOfLawItemAspect
     rdf:resource="ojns:JOL_2015_026_R_0001.ENG.pdfa1a.
                 l_02620150131en00010002.pdf"/>
 <cdm:hasSourceOfLawItemAspect
     rdf:resource="ojns:JOL_2015_026_R_0001.FRA.pdfa1a.
                 l_02620150131fr00010002.pdf"/>
</rdf:Description>
```

Moreover, the SourceOfLaw in the previous example has metadata (properties) related to its corresponding FRBR aspects, as well as the metadata of the FRBR aspects of its superclasses. An excerpt of its metadata at its Work level is the following:

```
<rdf:Description rdf:about="ojns:JOL_2015_026_R_0001">
 <rdf:type rdf:resource="cdm:SourceOfLawWork"/>
 <cdm:naturalIdentifier>
    L 26/1
 </cdm:naturalIdentifier>
```

[11] where `cellar:` represents the namespace `http://publications.europa.eu/resource/cellar/`

[12] where `cellar-record:` represents the namespace `http://publications.europa.eu/resource/cellar-record/`

```
 <cdm:issuingDate rdf:datatype="&xsd;dateTime">
    2015-01-26T00:00:00
 </cdm:issuingDate>
 <cdm:issuingAuthority>
    Council of the European Union
 </cdm:issuingAuthority>
</rdf:Description>
```

The described approach has been implemented as proof of concepts in RDF(S)/ OWL, resulting in the OWL-DL profile, thus available for deriving inferences by using DL reasoners like Pellet[13] or HermiT[14].

5 Benefits of the approach

The proposed CDM modeling approach has several advantages with respect to the existing one.

First of all it allows a direct constant access to the FRBR levels through the properties cdm:has[ResourceTypeFrbrClass]Aspect, while in the existing CDM the FRBR levels have to be navigated until reaching the expected one. In the existing CDM in fact there is no resource in the metadata that identifies the actual bibliographic entity (SourceOfLaw), therefore the resource can be either a work, expression, manifestation or item. Consequently a complex property path is necessary to navigate to the suitable FRBR entity and, in order to access to the Expression of a SourceOfLaw, for example, the following query is needed:

```
SELECT ?uri  WHERE
{
   ?resource cdm:item_belongs_to_manifestation?/
   cdm:manifestation_manifests_expression?/
   cdm:expression_belongs_to_work?/
   ^cdm:expression_belongs_to_work ?uri
}
```

On the contrary, in the new model the same result can be obtained by the following, more simple query:

```
SELECT ?uri WHERE
{
   ?resource rdf:type cdm:SourceOfLaw .
   ?resource cdm:hasSourceOfLawExpressionAspect ?uri
}
```

A similar query can be created to access all the FRBR aspects of an OP resource.

Another important advantage of this architecture is that the query for retrieving metadata of a resource is independent of its resource type. In fact, the inheritance mechanism on properties allows us to express queries at the top level of the hierarchy, independently of the resource type, while in the existing model

[13] http://clarkparsia.com/pellet/
[14] http://hermit-reasoner.com

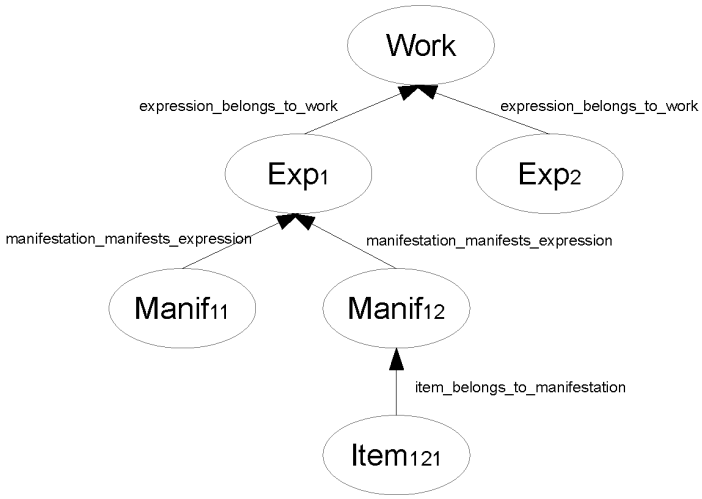

Fig. 5: Evaluating property paths in a FRBR hierachy

the resource type is to be known to retrieve its metadata, having as many queries as resource types.
For example, in the existing model, the following query can be run for accessing metadata at the level of Work of a SourceOfLaw:

```
SELECT * WHERE
{
   <RESOURCE_URI> cdm:item_belongs_to_manifestation?/
   cdm:manifestation_manifests_expression?/
   cdm:expression_belongs_to_work? ?uri .
   ?uri ?p ?o
}
```

The query retrieves all triples related to the work resource of a FRBR hierarchy. Regardless of the level of the entry point `<RESOURCE_URI>`, a property path leading to the root of the hierarchy is computed. The computation complexity obviously depends on the length of the path that needs to be evaluated [4]. Thus, processing a query starting at the Item level (path length 3) implies higher complexity than a query starting at the Work level (path length 0) as shown in Fig. 5.
In contrast, the new model enables clients to access any FRBR level using path expressions with a constant length of 1. For instance, the following query returns the Work aspect of a resource given the SourceOfLaw URI considered in Section 4, e.g., `cellar-record:58da3a99-a91d-11e4-8e01-01aa75ed71a1`:

```
SELECT DISTINCT ?uri WHERE
{
   cellar-record:58da3a99-a91d-11e4-8e01-01aa75ed71a1
   cdm:hasWorkAspect ?uri
```

```
}
```

Direct access to all FRBR levels via cdm:has[ResourceTypeFrbrClass]Aspect entirely eliminates the need for property paths. As a result, queries are simplified and evaluation complexity is reduced.
Note that this query does not contain any reference to the OP resource type. It can easily be adapted to any type of resource by changing the relation name, thus further simplifying the system's query framework.
For any resource URI representing the complete record, the task of accessing its metadata at different levels of the FRBR hierarchy becomes a matter of accessing levels of FRBR abstractions at the top of the resource taxonomy.
As another example, the subsequent query retrieves the English Expression of a resource in the current model:

```
SELECT ?uri WHERE
{
   <RESOURCE_URI> cdm: item_belongs_to_manifestation?/
   cdm:manifestation_manifests_expression?/
   cdm:expression_belongs_to_work?/
   ^cdm:expression_belo ngs_to_work ?uri.
   ?uri cdm:language "en"^^xsd:string.
}
```

Again, the query accounts for arbitrary starting points in the FRBR hierarchy, resulting increased complexity during evaluation.
In the new model, however, the same result is obtained by directly accessing the corresponding Expression aspect of the resource:

```
SELECT DISTINCT ?uri WHERE
{
   cellar-record:58da3a99-a91d-11e4-8e01-01aa75ed71a1
   cdm:hasExpressionAspect ?uri .
   ?uri cdm:language "en"^^xsd:string
}
```

Also in this case no reference to OP resource type is contained in the query.
An additional advantage of this modeling approach is the possibility to obtain a simplified management of the resource metadata, since they are organized in terms of properties of the FRBR classes, distributed at different levels of the resource taxonomy. This allows us, for example, to query the CDM model asking for all the Work metadata (i.e. owl:DatatypeProperties) of a generic SourceOfLaw, as follows:

```
SELECT DISTINCT ?property WHERE
{
   ?property rdf:type owl:DatatypeProperty .
   ?property rdfs:domain ?class .
   cdm:SourceOfLawWork rdfs:subClassOf* ?class
}
```

or to query the CDM model just selecting the specific metadata at SourceOfLaw-Work level:

```
SELECT DISTINCT ?property WHERE
{
   ?property rdf:type owl:DatatypeProperty .
   ?property rdfs:domain cdm:SourceOfLawWork
}
```

Secondly, in future versions of CELLAR, the new non-information resource that represents the actual bibliographic entity can be used as the proxy resource for aggregated views which contain metadata from all FRBR levels thus resolving the asymmetry of the content negotiation.

6 Conclusions

CELLAR represents the central information system of the OP, providing storage as well as advanced semantic indexing and access facilities to all the dissemination portals. The CDM semantic approach for the CELLAR resources is able to greatly improve accessibility of the OP multilingual documents. The proposed revision of the current CDM architecture, in particular, has the benefit of providing modularity and flexibility to the CDM approach, thus facilitating the management and extension of such knowledge organization system, as well as to simplify the query framework.

References

1. Study group on IFLA. Functional requirements for bibliographic records. Technical report, International Federation of Library Associations and Institutions, 1998. http://www.ifla.org/VII/s13/frbr/frbr.pdf.
2. Bianchini, C., Willer, M., ISBD resource and its description in the context of the semantic web. *Cataloging & Classification Quarterly*, 52:869–887, 2014.
3. Dunsire, G., Resource and work, expression, manifestation, item. Amended October 6, 2013, following comments by Patrick Le Boeuf and discussion at IFLA 2013, July 28 2013.
4. Katja Losemann and Wim Martens. 2013. The complexity of regular expressions and property paths in SPARQL. ACM Trans. Database Syst. 38, 4, Article 24 (December 2013), 39 pages.
5. R. Fielding and J. Reschke. Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content. Internet Engineering Task Force (IETF), RFC 7231, June 2014.
6. Architecture of the World Wide Web, Volume One. Ian Jacobs, Norman Walsh, Editors. World Wide Web Consortium, 15 December 2004.
7. Cool URIs for the Semantic Web. Leo Sauermann, Richard Cyganiak, Editors. World Wide Web Consortium, 3 December 2008.
8. SPARQL 1.1 Query Language. Steve Harris, Andy Seaborne, Editors. World Wide Web Consortium, 21 March 2013.