A Novel Vision for Navigation and Enrichment in Cultural Heritage Collections*

Joffrey Decourselle[1], Audun Vennesland[2], Trond Aalberg[2], Fabien Duchateau[1], and Nicolas Lumineau[1]

[1] LIRIS, UMR5205, Université Claude Bernard Lyon 1, Lyon, France
firstname.lastname@liris.cnrs.fr
[2] Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
firstname.lastname@idi.ntnu.no

Abstract. In the cultural heritage domain, there is a huge interest in utilizing semantic web technology and build services enabling users to query, explore and access the vast body of cultural heritage information that has been created over decades by memory institutions. For successful conversion of existing data into semantic web data, however, there is often a need to enhance and enrich the legacy data to validate and align it with other resources and reveal its full potential. In this visionary paper, we describe a framework for semantic enrichment that relies on the creation of thematic knowledge bases, i.e., about a given topic. These knowledge bases aggregate information by exploiting structured resources (e.g., Linked Open Data cloud) and by extracting new relationships from streams (e.g., Twitter) and textual documents (e.g., web pages). Our focused application in this paper is how this approach can be utilized when transforming library records into semantic web data based on the FRBR model in the process that commonly is called FRBRization.

Keywords: Cultural Heritage, Data Integration, Semantic Web, Linked Open Data, Entity Linking, Ontology and Entity Matching

1 Introduction

The last decade has seen a significant effort towards the use of semantic web data and related technologies. Linked Open Data (LOD) can be seen as "the Semantic Web done right" according to Tim Berners-Lee, with hundreds of interconnected knowledge bases (KBs) containing structured and semantic data [3]. However, the creation of reusable Linked Data from legacy data such as library records requires more than the transformation into new formats. The data often has to be transformed into acknowledged models (or type vocabularies) and needs to be correctly aligned with other resources before it appears as linked data. In the cultural heritage domain, the model in the Functional

* This work has been partially supported by the French Agency ANRT (*http://www.anrt.asso.fr*), the company PROGILONE (*http://www.progilone.fr*), a PHC Aurora funding (#34047VH) and a CNRS PICS funding (#PICS06945).

Requirements for Bibliographic Records (FRBR) [14] aims at representing data from cultural institutions with clear semantics [20], and it also offers benefits for improving search and visualization and new possibilities for semantic enrichment of cultural entities [8, 5]. To be widely adopted in cultural institutions, the FRBR model must be accompanied with a transformation process for converting legacy MARC data. The potential of FRBR lies in the relationships between entities, which unfortunately are rarely available in existing catalogues. Thus, it is necessary to enrich FRBRized collections with additional information from external data sources. Some relevant sources are already available on the Semantic Web, but a vast body of knowledge is still only available as text in documents (e.g., web pages). To facilitate the enrichment task needed in the FRBRization and other enrichment processes, the LOD and unstructured documents can be exploited. For instance, consider the Norwegian writer Henrik Ibsen. General information about this author are stored in knowledge bases such as DBpedia, VIAF or Freebase, uncommon facts are spread in fans web pages, and news about exhibitions related to his works might be available on streaming media such as Twitter. To provide a complete view of Ibsen's artistic life, it is necessary to aggregate this complementary, inconsistent and/or redundant knowledge from multiple heterogeneous data sources.

In this paper, we propose a generic framework for enriching FRBR collections. Our vision is to create thematic knowledge bases (TKBs) which gather relevant, reliable, and fresh information about a cultural topic (e.g., an artist, a work). The main objective of these TKBs is to help end-users and librarians discovering new knowledge. To build these TKBs, the idea is to exploit both types of data sources: the LOD, which is simpler to browse due to semantics but limited in terms of content, and the Web, with large amount of information but rather difficult to extract and with variable quality. Our framework aims at organizing the different processes involved in the building of a TKB, which are related to the following research areas: entity linking, information extraction, ontology/schema matching and entity matching. In addition, we explain how these processes should be adapted in the context of the cultural heritage domain, and we demonstrate the benefits of the TKB by presenting a use case.

In the rest of this paper, we first describe related work in Section 2. Then, Section 3 provides details about our framework for building thematic knowledge bases. Next, we illustrate the use of our framework with an enrichment scenario about *Natalie Dessay* (Section 4). We conclude by outlining future work.

2 Related Work

This work is at the crossroads of four research domains, namely entity linking, information extraction, and ontology and entity matching. We briefly present each of them in this section, and we also describe related projects.

Entity linking is the task of finding the corresponding entity (in a knowledge base) for a given mention (i.e., words used for labelling the entity). For instance, when the term *Tolkien* is found in a document, the objective is to decide whether

this refers to *dbpedia:J._R._R._Tolkien* or to *dbpedia:Christopher_Tolkien*. Due to the emergence of knowledge bases, that enable a long-term disambiguation, the named entity recognition community moved to entity linking [7, 11]. In our context, this task could be adapted to take into account the FRBR entity (from which the mention is extracted) and the application domain, which constrains the search of the corresponding entity in the knowledge base to the subset of entities related to cultural heritage [27].

Information extraction deals with the extraction of facts (i.e., relationships between two entities) from textual documents. Many issues arise since this task is at the crossroads of various research domains such as natural language processing, named entity recognition and data integration [26, 22]. Typical problems consist of entity linking (i.e., detecting and disambiguating entities based on their textual mentions and surrounding sentences), and discovering the type of relationship holding between two entities (e.g., by using generic patterns to represent sentences). Many approaches are able to extract facts (usually triples) from documents by considering the quality aspect (i.e., extraction of true facts) and the performance aspect (i.e., processing a large set of documents) [23, 28, 6]. A significant difference deals with the type of extraction: open information extraction means that new relationships can be created while the "closed" paradigm is limited to set of predefined relationships [9]. In our context, an open solution seems more interesting from the user point of view but more difficult to implement. Existing relationships between FRBR entities may be exploited to learn patterns rather than building them from textual documents.

Schema and ontology matching aims at solving the heterogeneity issues of data sources at the schema/ontology level by discovering semantic links (i.e., correspondences). These research fields have been largely studied in the literature [1, 10]. Traditionally, all possible pairs of elements are compared using similarity measures (e.g., Levenhstein distance), and the selection of the correspondences is performed using a decision maker (e.g., a threshold). As illustrated by the most recent challenges of Ontology Alignment Evaluation Initiative[3], traditional matching approaches cannot improve quality results any more. Thus, the new trend is to rely on complementary information, either from instances [16, 25, 18] or from user interactions. Similarly, we benefit from user feedback and these validations, as well as reuse of existing correspondences, need to be smartly integrated in the ontology matching process.

Entity matching, also known as record linkage, deals with the discovery of corresponding elements at the instance level, for example entities or records [17, 15]. The comparison of pairs (of elements) can be performed in a similar fashion as in the schema/ontology matching. However, the large amount of instances mainly requires a pre-processing step named blocking. Elements that share some common values (for a subset of their properties) are placed in the same block, and the comparison of pairs is applied inside every block in order to improve performance [4]. In our context, the selection of the best blocking key may be chosen according to statistics applied to the FRBRized collection. We may also

[3] `http://oaei.ontologymatching.org/`

benefit from the interconnections of LOD knowledge bases, which share common properties and may already link corresponding entities.

Related projects. The previously described research domains mainly focus on a single issue, and propose generic solutions to solve them. In the context of semantic enrichment for cultural heritage, we need to tackle the same issues but each solution can be adapted to benefit from the FRBR model. Europeana[4] is the closest project to our work and it shares some common goals in terms of enrichment [13]. However, it aims at creating a centralized authoritative source while we believe that cultural institutions should be responsible for managing their resources. In Europeana, a first proposition for enrichment is based on machine-learning algorithms to extract relevant added values [2]. Later, an automatic enrichment is proposed, but limited to four properties, for instance places (with links to Geonames) or agents (with links to DBPedia). Since the enrichment is performed at large scale, the frequency of errors can only be estimated: it reaches 1.5% of the dataset, which still represents more than 15,000 errors [24]. In our proposition, we combine user interactions and reuse of validated knowledge to favour a high quality enrichment. Besides, our work is one step beyond by proposing the integration of textual and streaming contents.

3 Framework for Building Thematic Knowledge Bases

In this section, we introduce our framework for building thematic knowledge bases (TKBs). Note that the model of the TKB is out of scope of this paper, but we expect it as open as possible according to user requirements. Indeed, the representation of basic properties for the cultural heritage entities is covered in the FRBR specifications. For additional information, it is either possible to use existing ontologies (e.g., Linked Open Vocabularies[5]) or to create a specific one. As illustrated in Figure 1, our framework includes four main processes (square boxes) and uses as input a mention (e.g., the title of a FRBR Work, the name of an Agent). Each process can be seen as a black box, and we describe each of them in the rest of this section.

3.1 Entity linking

In the cultural heritage domain, many artistic works can be found on the LOD. The exploitation of this resource at the first place is therefore relevant. The entity linking process uses a mention as input and it detects the LOD entities related to this mention (one entity per LOD knowledge base). Contrary to existing entity linking approaches [27, 7, 11], this process has two specificities in our framework. First, entity linking approaches traditionally use a mention with surrounding terms (e.g., sentence) while our input mention is part of a FRBR entity (mainly a Work or an Agent). Secondly, we are not limited to search on a single LOD

[4] `http://www.europeana.eu/`

[5] `http://lov.okfn.org/dataset/lov/`

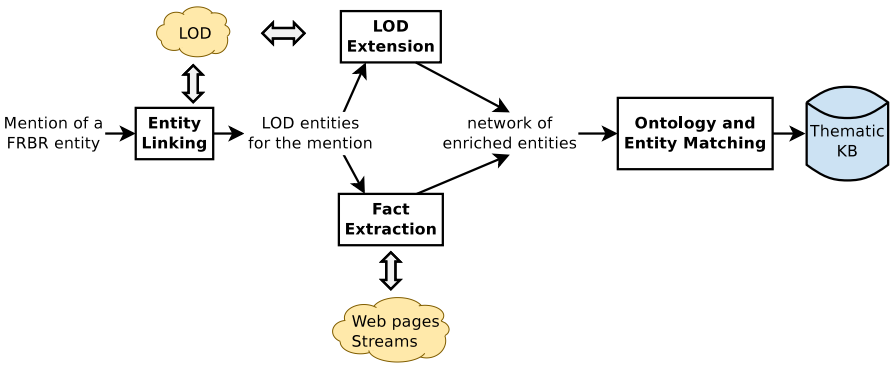

Fig. 1. Framework for building thematic knowledge bases

knowledge base. This means that we can exploit the interconnection between LOD knowledge bases (e.g., *owl:sameAs* predicates) to improve the accuracy of the entity linking [12]. For instance, linking the mention *"Henrik Ibsen"* to both LOD entities *dbpedia:Henrik_Ibsen* and *freebase:/m/03pm9* reinforces the confidence that the linking is correct since the two LOD entities are connected through a property *owl:sameAs*. The discovered LOD entities are then processed in parallel by the LOD extension and the fact extraction.

3.2 LOD extension

The LOD extension process aims at discovering new entities related to the LOD entities which represent the mention. As LOD knowledge bases are reputed for their good quality, the goal of this process is to integrate reliable information in the TKB. The main challenge lies in the level of extension. LOD knowledge bases, especially general ones, may contain facts that are either too broad or not useful for semantic enrichment or user navigation. Similarly, some properties can include a long list of values (e.g., *owl:sameAs*, *rdf:type*) and should be filtered to avoid overloading the TKB. A possible solution for the LOD extension could be an iterative process in which a limited number of facts is added and evaluated at each iteration until the result is satisfying, for instance in terms of consistency. Such process should also take into account the FRBR context. The relationships between FRBR entities and the FRBR attributes can be associated to LOD properties. Back to our running example, the Ibsen DBpedia entity enables us to extend to *dbpedia:The_Wild_Duck*, one of his play, or to *dbpedia:August_Strindberg*, another novelist who influenced Ibsen. With LOD extension, we gather new facts about our initial mention. Note that some facts may be redundant (e.g., provided by two KBs) and the cleaning is performed in the ontology/entity matching process.

3.3 Fact extraction

In addition to the LOD extension, our framework exploits unstructured documents with variable quality, such as web pages. The main motivation is to gather uncommon facts (i.e., that are not present in the LOD). In our context, the initial mention and the LOD entities obtained from the entity linking step are used to identify interesting documents and detect sentences about the entity. To enable enrichment in the TKB, new relationships and properties can be added to the FRBR model, thus promoting the use of an open information extraction tool (see Section 2). Another notable difference deals with human intervention: cataloguers who validate a discovered fact for enriching a TKB also provides feedback for future fact extraction. This means that the sentence (or pattern) can be stored and marked as reliable, the predicate of the new fact is validated, etc. Finally, microblogging data sources such as Twitter contain abbreviations, notations (e.g., hashtags) which require specific solutions. The FRBR collection is helpful for learning patterns: if an Agent entity does not have a given property that other Agent entities own, we can learn the patterns or sentences in which this fact could be detected. In our running example, we could extract from this blog[6] the fact <*dbpedia:Henrik_Ibsen*, *relationship:acquaintanceOf*, *dbpedia:James_Joyce*>, where two DBpedia entities representing famous novelists are linked through the *acquaintanceOf* property (from the *relationship* vocabulary[7]). At the end of this task, the initial LOD entities have been enriched with new facts, and new entities may have been created. This additional knowledge can include redundancies, that the next task is in charge of cleaning.

3.4 Ontology and Entity matching

In this last process, we perform ontology and entity matching to clean extracted information. In the same LOD entity, different properties can represent the same concept, which in turn may be redundant with a FRBR attribute (e.g., the *name* of a FRBR agent is equivalent to the LOD properties *dbpprop:name* and *foaf:name*). As our framework is more suited to an open extraction approach, it produces facts whose predicate needs to be mapped to an existing one to avoid redundancies. Solving these issues requires both ontology matching and entity matching, two research areas traditionally considered separately [1, 10]. In our framework, the idea is to combine ontology and entity matching in order to clean the "network" of enriched entities resulting from the previous steps. Correspondences at the ontology level are needed to perform entity matching. Conversely, we believe that entity matching results can reinforce the discovery of new correspondences at the ontology level. A basic ontology matching approach can be used as a bootstrap process to detect correspondences between concepts and properties, then entity matching is performed to discover entity correspondences. An iterative combination of both matching processes enables

[6] `http://blog.bookstellyouwhy.com/archive/2015/03` (article from March 18^{th})
[7] `http://vocab.org/relationship/`

the refinement of existing correspondences and the discovery of more complex ones. When possible, extracted relationships have to be mapped to FRBR relationships or attributes by relying on ontology matching too. In case of conflicting correspondences, information provenance may be useful since LOD extension is considered as more reliable than fact extraction. Note that the correspondences at the ontology level needs to be stored for reuse. Data fusion (a.k.a. augmentation), which consists in merging redundant or complementary information, is optional in our context. The cultural heritage expert may decide how to select or merge information from the TKB. In the Ibsen example, we could obtain during LOD extension the entities *dbpedia:The_Wild_Duck* and *viaf:312333678*, which both represent the novel *Wild Duck*. By applying an entity matching process, a correspondence between the two entities is discovered and only one of them is added to the TKB (and a property *sameAs* is used to link to the discarded entity). At the end of the matching process, the TKB is built and ready for use, as illustrated in the next section with a real-world scenario.

4 Building a TKB about *Natalie Dessay*

A TKB is useful both for experts who need to enrich their original collections and for library users who can explore it for finding new resources related to their initial query. Since an implementation of our framework is currently in progress, this section explains how it can be applied in a semantic enrichment scenario.
Let us describe an example about the well-known French singer *Natalie Dessay*. A librarian has generated a FRBR entity for the Agent *Natalie Dessay*, but there is almost no information in the original records about the singer. Thus she decides to build a TKB to enrich the FRBR entity. As shown in Figure 2, general information about *Natalie Dessay* are stored in LOD knowledge bases such as DBpedia[8] or Freebase[9], uncommon information could be found in textual websites[10], and news about her activities (e.g., concerts, TV appearance) might be available on streaming media such as Twitter[11].
Instead of manually querying and browsing these multiple data sources, the librarian simply runs an implementation of our framework with the input mention *Natalie Dessay* and the corresponding FRBR Agent entity. During the first step, the mention is linked to its corresponding entities in DBpedia and Freebase (respectively *dbpedia:Natalie_Dessay* and *fb:m.0cfmsz*). The search engines of both general knowledge bases ranks the correct entity at the top. Since both LOD entities are already linked through a *owl:sameAs* predicate, it increases the confidence in this discovery. The LOD extension consists in gathering facts from the LOD entities, e.g., the triples <*dbpedia:Natalie_Dessay*, *dbpprop:birthDate*, *'1965-04-19'*> and <*fb:m.0cfmsz*, *fb:people.person.date_of_birth*, *'1965-04-19'*>. To avoid overloading the TKB, mainly with properties which have numerous

[8] *dbpedia:Natalie_Dessay*, `http://dbpedia.org/page/Natalie_Dessay`
[9] *fb:/m/0cfmsz*, `https://www.freebase.com/m/0cfmsz`
[10] `http://blogclarabel.canalblog.com/archives/2014/12/09/31081571.html`
[11] `https://twitter.com/n2cfan`

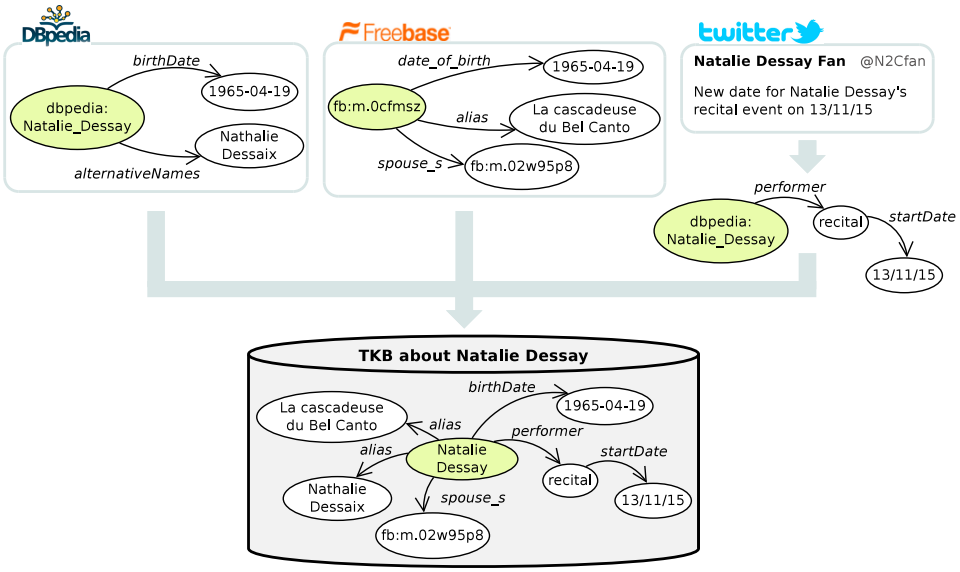

Fig. 2. From data sources to the thematic knowledge base about *Natalie Dessay*

values, we decide to apply a filter based on the provenance and to only store properties and values from the *DBpedia* and the *Freebase* ontologies. The LOD extension also provides alternatives mentions using alias properties such as *dbpprop:alternativeNames* or *fb:common.topic.alias*, thus enabling us to collect the extra mention *Nathalie Dessaix*. These alternatives mentions are exploited during the third process, fact extraction. In the documents containing the mentions, we discover a tweet about an upcoming recital of Nathalie Dessay. As shown in Figure 2, information extraction tools are needed to transform natural language into triples. Note that new predicates such as *performer* can either be available in the FRBR model or based on an external ontology (e.g., *schema.org*). In the ontology and entity matching step, cleaning is performed to remove redundancies or to reduce heterogeneity in the TKB. Mappings are performed at the ontology level (e.g., between the properties *dbpprop:birthDate* and *fb:people.person.date_of_birth*), or at the entity level (e.g., to check that *Laurent Naouri*, the husband of Natalie Dessay, is not represented by two entities in the TKB). Mappings can be validated by the user and reused later, specifically those at the ontology level. When the TKB about *Natalie Dessay* is constructed, the librarian can select the relevant facts to enrich the FRBR initial entity.

5 Conclusion

This paper introduces a new framework for enriching cultural heritage collections. It is based on the notion of thematic knowledge bases, which aggregates

information about a given topic from various sources. Our framework for building these TKBs includes generic processes, but we have explained how the specificities of FRBR may enhance each process. Finally, a scenario illustrating the semantic enrichment has validated our approach.
The first perspective is to implement and experiment the framework, mainly in terms of quality. A TKB has to find a tradeoff between accuracy (i.e., a high rate of true facts) and completeness (i.e., a high rate of new facts which are not initially covered by the FRBR entities). The usability of these TKBs, and graphical solutions to browse them, needs to be tested, for instance when representing complex bibliographic relationships [21]. The selection of relevant data sources is also an interesting challenge (e.g., VIAF is useful for writers while MusicBrainz is more appropriate for musicians). Another motivation is to extend these TKBs as a new search paradigm. Searching for information is still performed like twenty years ago, by querying a search engine and browsing documents. Similarly to aggregated search [19], our TKB could go further by involving new challenges for gardening, indexing or sharing knowledge.

References

1. Bellahsene, Z., Bonifati, A., Rahm, E. (eds.): Schema Matching and Mapping. Data-Centric Systems and Applications, Springer (2011)
2. Berardi, G., Esuli, A., Gordea, S., Marcheggiani, D., Sebastiani, F.: Metadata enrichment services for the europeana digital library. In: Theory and Practice of Digital Libraries. pp. 508–511. TPDL'12, Springer (2012)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
4. Böhm, C., de Melo, G., Naumann, F., Weikum, G.: Linda: Distributed web-of-data-scale entity matching. In: International Conference on Information and Knowledge Management. pp. 2104–2108. CIKM '12, ACM (2012)
5. Buchanan, G.: FRBR: Enriching and integrating digital libraries. In: JCDL (June 2006)
6. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.R.H., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Proceedings of Artificial Intelligence (AAAI 2010) (2010)
7. Dai, H.J., Wu, C.Y., Tsai, R., Hsu, W.L.: From entity recognition to entity linking: a survey of advanced entity linking techniques. In: The 26th Annual Conference of the Japanese Society for Artificial Intelligence. pp. 1–10 (2012)
8. Dickey, T.J.: FRBRization of a Library Catalog: Better Collocation of Records, Leading to Enhanced Search, Retrieval, and Display. Information Technology & Libraries 27, 23–32 (2008)
9. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communication of ACM 51, 68–74 (Dec 2008)
10. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, 2nd edn. (2013)
11. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with wikipedia. Artificial intelligence 194, 130–150 (2013)
12. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameas isnt the same: An analysis of identity in linked data. In: The Semantic Web–ISWC 2010, pp. 305–320. Springer Berlin Heidelberg (2010)

13. Haslhofer, B., Momeni, E., Gay, M., Simon, R.: Augmenting europeana content with linked data resources. In: Proceedings of the 6th International Conference on Semantic Systems. pp. 40:1–40:3. I-SEMANTICS '10, ACM (2010)
14. IFLA: Functional requirements for bibliographic records: final report. Tech. rep., IFLA (February 2009)
15. Ioannou, E., Rassadko, N., Velegrakis, Y.: On Generating Benchmark Data for Entity Matching. Journal on Data Semantics 2(1), 37–56 (2013)
16. Isaac, A., van der Meij, L., Schlobach, S., Wang, S.: An empirical study of instance-based ontology matching. In: The Semantic Web, Lecture Notes in Computer Science, vol. 4825, pp. 253–266. Springer (2007)
17. Köpcke, H., Thor, A., Rahm, E.: Evaluation of entity resolution approaches on real-world match problems. PVLDB 3(1), 484–493 (2010)
18. Lacoste-Julien, S., Palla, K., Davies, A., Kasneci, G., Graepel, T., Ghahramani, Z.: Sigma: Simple greedy matching for aligning large knowledge bases. In: International Conference on Knowledge Discovery and Data Mining. pp. 572–580. KDD '13, ACM (2013)
19. Lalmas, M.: Aggregated search. In: Melucci, M., Baeza-Yates, R. (eds.) Advanced Topics in Information Retrieval, The Information Retrieval Series, vol. 33, pp. 109–123. Springer (2011)
20. Le Bœuf, P.: FRBR: Hype or cure-all? Introduction. Cataloging & Classification Quarterly 39(3-), 1–13 (2005)
21. Mercun, T., Zumer, M., Aalberg, T.: Presenting and exploring the complexity of bibliographic relationships. In: ICADL. pp. 63–66 (2012)
22. Nakashole, N., Weikum, G., Suchanek, F.M.: Discovering and exploring relations on the web. PVLDB 5(12), 1982–1985 (2012)
23. Parameswaran, A., Garcia-Molina, H., Rajaraman, A.: Towards the web of concepts: extracting concepts from large datasets. Proceedings of VLDB Endowment 3, 566–577 (2010)
24. Stiller, J., Petras, V., Gde, M., Isaac, A.: Automatic enrichments with controlled vocabularies in europeana: Challenges and consequences. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection, Lecture Notes in Computer Science, vol. 8740, pp. 238–247. Springer (2014)
25. Suchanek, F.M., Abiteboul, S., Senellart, P.: Paris: Probabilistic alignment of relations, instances, and schema. Proc. VLDB Endow. 5(3), 157–168 (Nov 2011)
26. Suchanek, F.M., Sozio, M., Weikum, G.: SOFIE: A Self-Organizing Framework for Information Extraction. In: International World Wide Web conference (WWW 2009). ACM (2009)
27. Takhirov, N., Duchateau, F., Aalberg, T.: Linking FRBR Entities to LOD through Semantic Matching. In: Theory and Practice of Digital Libraries (TPDL). pp. 284–295. Springer (2011)
28. Takhirov, N., Duchateau, F., Aalberg, T.: An Evidence-based Verification Approach to Extract Entities for Knowledge Base Population. In: International Semantic Web Conference (ISWC). pp. 575–590. Springer (2012)