Educational Linked Data on the Web - Exploring and Analysing the Scope and Coverage

Davide Taibi[1(✉)], Giovanni Fulantelli[1], Stefan Dietze[2], and Besnik Fetahu[2]

[1] Istituto per le Tecnologie Didattiche, Consiglio Nazionale delle Ricerche, Palermo, Italy
{davide.taibi,giovanni.fulantelli}@itd.cnr.it
[2] L3S Research Center, Hannover, Germany
{dietze,fetahu}@l3s.de

Abstract. Throughout the last few years, the scale and diversity of datasets published according to Linked Data (LD) principles has increased and also led to the emergence of a wide range of data of educational relevance. However, sufficient insights into the state, coverage and scope of available educational Linked Data seem still missing. In this work, we analyse the scope and coverage of educational linked data on the Web, identifying the most significant resource types and topics and apparent gaps. As part of our findings, results indicate a prevalent bias towards data in areas such as the life sciences as well as computing-related topics. In addition, we investigate the strong correlation of resource types and topics, where specific types have a tendency to be associated with particular types of categories, i.e. topics. Given this correlation, we argue that a dataset is best understood when considering its topics, in the context of its specific resource types. Based on this finding, we also present a Web data exploration tool, which builds on these findings and allows users to navigate through educational linked datasets by considering specific type and topic combinations.

Keywords: Dataset profile · Linked data for education · Linked data explorer

1 Introduction

The diversity of datasets published according to Linked Data (LD) [5–7] principles has increased in the last few years and also led to the emergence of a wide range of data of educational relevance [18]. These include open educational resources metadata, statistical data about the educational sector, video lecture metadata or university data about courses, research or experts [2]. Initial efforts to collect and catalogue such datasets have been made through initiatives such as the LinkedUp Data Catalog[1] or related community initiatives[2].

[1] http://data.linkededucation.org/linkedup/catalog/.

[2] These include, for instance, http://linkededucation.org, http://linkeduniversities.org or the recently established W3C Community Group on Open Linked Education (http://www.w3.org/community/opened/).

D. Mouromtsev and M. d'Aquin (Eds.): Open Data for Education, LNCS 9500, pp. 16–37, 2016.
DOI: 10.1007/978-3-319-30493-9_2

However, the state, coverage and scope of available educational Linked Data have not been widely investigated. Here, in particular questions about the represented resource types, such as, resource metadata or information about organisations or people, and topics are of crucial relevance to shape a better understanding about the state of educationally relevant Linked Data on the Web. Also identifying a dataset containing resources related to a specific topic is, at present, a challenging activity. Moreover, the lack of up-to-date and precise descriptive information has exacerbated this challenge. The mere keywords-based classification derived from the description of the dataset owner is not sufficient, and for this reason, it is necessary to find new methods that exploit the characteristics of the resources within the datasets to provide useful hints about topics covered by datasets and their subsequent classification.

In this direction, authors in [1, 3] proposed an approach to create structured metadata to describe a dataset by means of topics, defined as DBpedia categories, where a weighted graph of topics constitutes a dataset profile. Profiles are created by means of a processing pipeline[3] that combines techniques for datasets resource sampling, topic extraction and topic ranking. Topics have been extracted by using named entity recognition (NER) techniques, where topics are ranked, respectively weighted, according to their relevance using graph-based algorithms such as PageRank, K-Step Markov, and HITS.

The limitations of such an approach are related mainly to the following aspects. First, the meaning of individual topics assigned to a dataset can be highly dependent on the type of resources they are attached to. Also, the entire topic profile of a dataset is hard to interpret if categories from different types are considered at the same time. As an example of the first issue, the same category (e.g. “Technology”) might be associated to resources of very different types such as “video” (e.g. in the Yovisto Datset[4]) or “research institution”(e.g. in the CNR dataset[5]). Concerning the second issue, the single topic profile attached for instance to bibliographic datasets (such as: the LAK dataset[6] or Semantic Web Dog Food[7]) - in which people (“authors”), organisations (“affiliations”) and documents (“papers”) are represented– is characterized by the diversity of its categories (e.g. DBpedia categories: Scientific_disciplines, Data_management Information_science but also Universities_by_country, Universities_and_colleges). Indeed, classification of datasets in the LD Cloud is highly specific to the resource types one is looking at. While one might be interested in the classification of “persons” listed in one dataset (for instance, to learn more about the origin countries of authors in DBLP), another one might be interested in the classification of topics covered by the documents (for instance disciplines of scientific publications) in the very same dataset.

[3] http://data-observatory.org/lod-profiles/profiling.htm.

[4] http://www.yovisto.com/.

[5] http://data.cnr.it/.

[6] http://lak.linkededucation.org.

[7] http://data.semanticweb.org.

In this paper, we aim at providing a systematic assessment of educational Linked Data which consider both, represented topics as well as resource types and their correlations. Questions of interest are:

1. Which types and topics are covered by existing educational Linked Data?
2. What are the central topics covered for particular types (e.g. Open Educational Resources metadata)?
3. Are certain topics underrepresented for certain types, or vice versa?

The approach we propose overcomes the limitations described above by considering the topic profiles defined in [3] in the context of the resource types they are associated with. However, the schemas adopted by the datasets of the LD cloud are heterogeneous, thus making it difficult to compare the topic profiles across datasets. While there are many overlapping type definitions representing the same or similar real world entities, such as "documents", "people", "organization", type-specific profiling relies on type mappings to improve the comparability and interpretation of types and consequently, profiles. For this aim the explicit mappings and relations declared within specific schemas (for instance, *foaf:Person* being a subclass of *foaf:Agent*) as well as across schemas (for instance through *owl:equivalentClass* or *rdfs:subClassOf* properties) are crucial.

While relying on explicit type mappings, we have based our work on a set of datasets where explicit schema mappings are available from earlier work [2]. This includes education-related datasets identified by the LinkedUp Catalog[8] in combination with the dataset profiles generated by the Linked Data Observatory[9]. While the latter provides topic profiles for the majority of LOD datasets, the LinkedUp Catalog contains explicit schema mappings which were manually created for the most frequent types in the LinkedUp Catalog.

The next Section provides a broad overview on the educational Linked Data from a perspective that highlights the relations with the Open Educational Resource world; then, we provide a thorough state of the art assessment of the coverage and scope of educational Linked Data in Sect. 3, which answer aforementioned questions. In addition, we introduce an interactive explorer of educational Linked Data, in Sect. 4, which aims at providing a resource type-specific view on categories associated with the datasets in the LinkedUp Catalog.

2 Resources for Education: Linked Data and OER

The Semantic Web, and specifically the possibility to publish data on the Web and connect them through links (i.e. the Linked Data model), represents one of the most significant evolution of the Internet, after the idea of the Web itself.

[8] http://data.linkededucation.org/linkedup/catalog/.

[9] http://data-observatory.org/lod-profiles.

From an educational point of view, both the human-readable and navigable structure of the Web pages and the machine-processable datasets of LD have opened up incredible potentials for the implementation of new and effective pedagogical paradigms [4].

The hypertextual organization of information and knowledge of the Web has influenced not only the ICT-based educational projects worldwide, but also the publication of traditional school textbooks, where anchor-like notes appear throughout a book for immediate references to other related concepts.

In general, the more evident opportunity of the Web for education is a very basic one, yet extremely important for education: the possibility to publish information that everybody can access and use to develop knowledge.

Some years after the birth of the Web, under the pressure of economical, philanthropic and pedagogical emergencies, the idea to exploit materials published on the Web for educational purposes brought to the development of the Open Educational Resources (OER) movement. Since then, hundreds of OER repositories have populated the Web with resources designed for education.

From a pedagogical perspective, OER have solved some critics related to the use of Web pages for education, such as the lack of a pedagogical structure to present information, or the difficulties in identifying the pedagogical scope of a resource published on the Web. However, the OER movement does not exclude more general resources accessible through the Web, provided that they are included into usage patterns designed according to pedagogical criteria. Furthermore, the spectrum of OER is really wide, ranging from resources produced by academic or educational committed institutions to user- or crowd-generated resources. This variety of OER is reflected in the many definitions of OER that can be found in the literature [13–17].

In spite of their tremendous influence on education, OER have also shown some limits; amongst the others:

- The lack of a sole standard for OER and repositories, which has fragmented the offer of OER on the Web [8];
- The complexity in handling direct links between OER and, consequently, in finding semantically related resources.
- The impossibility to guarantee metadata interoperability, due to the proliferation of educational metadata schemas [9];
- The impossibility to deal with the vast availability of education-related data on the web.

The Linked Open Data model offers new solutions for educational resources, partly solving some of the OER limits, still representing a paradigm that complements the OER one, and does not substitute it.

Amongst the OER issues that can be solved by the LOD approach:

- LOD are interlinked by definition; consequently, algorithms can automatically identify semantically related resources; in the case of OER, it was necessary to develop a semantic layer to describe OERs;

- Federated query can be used in order to find resources belonging to distinct datasets; as far as OER are concerned, this was only possible if OER repositories were federated, e.g. through the OAI PMH which allows the exposition of metadata through a common protocol;
- LOD provide the solutions to publish education-related data on the web.

For these reasons, the interest of the educational community in LOD has developed over the years, even sustained by the growing availability of resources published in the Linked Data format, which has raised from 12 in 2007 up to 570 in 2014[10].

The first applications of Linked Data to education focused on the potentials of LD to solve interoperability issues in the field of TEL (Technology Enhanced Learning). In the mEducator project [12], data from a number of open TEL data repositories has been integrated, exposed and enriched by following LD principles. Afterwards, more and more attention has been paid to the increasing availability of datasets on the Web, and particularly to the presence of educational information in the linked data landscape.

The LinkedUp project has explicitly aimed at the educational exploitation of Linked Open Data, and has distinguished two types of linked datasets: datasets directly related to educational material and institutions, including information from open educational repositories and data produced by universities; datasets that can be used in teaching and learning scenarios, while not being directly published for this purpose.

Therefore, the approach followed by the LinkedUp project enhances the general principle of the OER movement that not only resources explicitly developed for educational purposes can be used in educational patterns.

From one hand, this is an essential advantage of open education in general; however, it amplifies some drawbacks that could hinder the potentials of LOD in education:

- *Which datasets and resources can be employed in educational contexts?* A similar challenge has been already addressed in the OER world. However, this task presents a higher degree of complexity for LOD, since the OER movement focuses on the development of content on pedagogical principles, while generally there is no pedagogical theory behind the publication of a dataset, and classifying them becomes more complicated.
- *How datasets (and their resources) should be described in order to facilitate their search (and pedagogical exploitation)?* This issue shows one of the main difference between OERs and LOD. While a bad-described OER can be easily visited by the end user in order to check if it is suitable for a specific educational project, a bad-described dataset can be hardly analysed by the end user, and the risk that the dataset will be ignored is very high.

For these reasons, specific classification mechanisms as the ones described in this chapter, which highlight the key elements of a dataset, together with search tools based on the, are extremely important to fully exploit the potentials of LOD in education.

[10] source: http://lod-cloud.net/.

3 Analysing the Coverage of Educational Linked Data

In this section, we present the actual analysis of educational Linked Datasets on the Web, taking into account both topics as well as resource types.

3.1 Data and Method

Topic annotations are provided in the form of DBpedia categories for the majority of LD datasets, available from the topic profiles[11] dataset, further described in [3]. A topic profile connects a dataset with the topics extracted from the analysis of resource samples. Since topics are ranked, a topic profile can be seen as a weighted dataset-topic graph. As such a, topic profile provides a comprehensive overview of the topic coverage of individual datasets. Analysed across a specific set of datasets - as carried out in this work - topic profiles provide insights into the coverage of such a set of datasets.

While topic annotations are obtained from analysing resources of a particular type, the semantics of the topic can best be interpreted when considering the type of the resource. As an example, if the topic "Biology" is associated to a resource of type *foaf: Document*, for instance, a scholarly publication, it indicates that this particular resource is related to biological aspects. In case the "Biology" topic is associated to a *foaf: Organization* resource, it is likely referred to a Biology department of a university. Next to such differences in interpreting topics, the nature of DBpedia categories also differs significantly across different types. For instance, while actual document-related types usually are related to topics which indicate some form of subject or domain (such as "Biology"), resources which represent some notion of organisation or person usually are characterised through some broader categorisations, such as "Academic_institutions" or "People_from_Athens". These fundamental differences are important to understand the nature of dataset topic profiles and to motivate our adopted methodology.

Since our work considers the investigation of both, topics and types, we use as additional data source the LinkedUp Catalog[8]. Our research investigates 21 datasets, which is precisely the set of datasets existing in both collections the LinkedUp Catalog and the Dataset Topic Profiles, as only for these both topic profile and resource type mapping annotations were available. The complete list of selected datasets is shown in Table 1. As explained by Fetahu et al. in [3], topic profiles are generated based on resource samples, where the applied sampling strategies did take into account factors such as the population size of respective types leading to different sample sizes across different datasets. Table 1 indicates both the total amount of data and the characteristics of the automatically computed sample.

The analysis of the relationships between datasets, topics and resource types - aimed at providing a response to the research questions posed above - has been undertaken exploiting network analysis theories and methods. Indeed, the connections between the three investigated notions can be represented by networks, in which

[11] http://data.l3s.de/dataset/linked-dataset-profiles.

Table 1. Datasets, resources and resource types

Dataset	Total data		Sampled data		
	#Types	#resource	# Types	#resource	# Categ.
asn-us	29	7494200	3	10000	2128
Colinda	21	17006	9	1985	479
data-cnr-it	120	485977	7	29768	2702
data-open-ac-uk	134	386291	7	36668	1979
education-data-gov-uk	99	315632	42	18712	2510
educationalprograms_sisvu	27	104238	22	12627	2144
gesis-thesoz	9	48532	4	1176	487
hud-library-usagedata	6	904747	1	10000	2300
l3s-dblp	6	15514	3	9368	943
lak-dataset	14	13688	3	10000	1496
linked-open-aalto-data-service	22	373553	12	17598	1543
Morelab	13	244	9	890	206
open-courseware-consortium-metadata-in-rdf	4	22850	4	29369	2723
organic-edunet	1	11093	1	847	559
Oxpoints	142	73655	30	8649	1529
publications-of-charles-university-in-prague	258	14324	15	658	197
seek-at-wd-ict-tools-for-education-web-share	556	13502	37	9938	1624
unistat-kis-in-rdf-key-information-set-uk-universities	35	371737	9	39684	556
universitat-pompeu-fabra-linked-data	39	5778	13	1617	312
university-of-bristol	15	240179	12	22572	2450
Yovisto	8	549986	8	5605	2122

the elements are nodes and their relationship are edges. Specifically the analysis of the relationships has been conducted by considering:

- the network representing the relationships between datasets mediated by categories/topics
- the network representing the relationships between datasets mediated by resource types
- the network representing the relationships between resource types mediated by categories/topics

These networks have been represented by using the Open Source software Gephi[12]. Due to the high number of categories connected to certain datasets (as shown in Table 1), dataset profiles have been filtered by selecting for each dataset the top 100 categories with the highest relevance score. Exploiting the insights gained from such networks, we can identify the particular type/topic coverage of educational LD datasets, corresponding gaps, and the correlation of educational resource types and topics.

3.2 Analysing Topic Coverage - the Dataset-Category-Graph

Representing datasets and categories, i.e. topics, as a weighted graph allows us to analyse the topic coverage of assessed datasets and their proximity topic-wise. In particular,

[12] http://gephi.github.io/.

a dataset is connected with the corresponding category depending on its topic profile. Indirect relationships among datasets emerge through shared or connected categories.

The nodes of this network represent datasets and categories. In particular, a dataset is connected with the correspondent category depending on its topic profile. For this reason there is not a direct connection between two datasets, but the categories act as indirect links by connecting datasets that share the same categories. Representing datasets and categories, i.e. topics, as a weighted graph allows us to analyse the topic coverage of assessed datasets and their proximity topic-wise. In particular, a dataset is connected with the corresponding category depending on its topic profile. Indirect relationships among datasets emerge through shared or connected categories.

The color and the dimension of the nodes are related to the metrics of the network which were calculated. In particular, the color gradient is related to the degree of a node (a darker node has a higher degree) while the dimension is proportional to the *betweenness centrality* measure. A detailed view of the graph is shown in Fig. 2. Next to other measures that indicate the importance of the nodes based on their topology, the betweenness centrality of a node is calculated by considering the number of the shortest paths from all pairs of nodes that pass through the node.

Table 2 reports the list of the top ten most connected categories in the datasets under investigation by taking into consideration the number of datasets. Note that while each topic is a DBpedia category, we omitted the DBpedia namespace (http://dbpedia.org/category/) from the listing. The number of datasets sharing the specific category is also reported.

Table 2. Top 10 categories according to their number of occurrences in distinct dataset profiles

Category	# dataset
Academic_disciplines	19
Applied_sciences	19
Applied_disciplines	18
Applied_mathematics	18
Artificial_intelligence	18
Areas_of_computer_science	16
Formal_sciences	16
Interdisciplinary_fields	16
Computing	16
Biology	16

Ranking categories according to the number of resources they are associated with, a different set of top-10 categories emerges (Table 3). The number of datasets sharing the specific category is also reported.

The categories reported in Table 3 highlight the heterogeneity of the dataset resources: categories representing actual disciplines (such as *Biology*, *Computing*, as extracted from Open Educational Resources or video lectures) as well as categories related to institutions (such as *Academic_institutions*) are represented in the list. This overview already demonstrates the strong impact of the resource type (eg *foaf:*

Table 3. Most represented categories in the sampled resources

Category	# dataset	# resources	#types
Applied_sciences	19	3581	81
Computing	16	2778	92
Academic_disciplines	19	2328	68
Biology	16	2068	56
Digital_technology	12	2012	51
Education	14	1855	66
Academia	15	1668	63
Academic_institutions	14	1625	54
Interdisciplinary_fields	16	1574	57
Society	12	1476	60

Document or *foaf:Organisation*) on the associated categories, an observation which motivated parts of the following investigations and an explorative browser described in [11] and Sect. 4.

In the network of Fig. 1 the resource type is not considered, thus two datasets can be connected even if they are collecting different types of resources such as information about institutions, learning materials or scientific publications. In this way, the description of each dataset is not only based on keywords and description provided by dataset authors, but useful hints are also provided about the topics to which the resources of a dataset are connected with. Moreover, the network allows to identify clusters of dataset containing resources related to similar topics.

In order to investigate furthermore the influence of the dataset heterogeneity, the following four datasets containing resources of different nature has been selected:

- The LAK Dataset[13] providing scholarly papers in the Educational Data Mining and Learning Analytics research fields [10].
- The data.cnr.it[14] providing information about the National Research Council of Italy institutes.
- The course descriptions contained in the Linked Data endpoint of The Open University UK[15].
- The L3S DBLP[16] dataset that collects papers related to computer science discipline.

All these datasets provide resources for different educational purposes. In fact, the CNR dataset describes the organization level of this institution, providing information about buildings and persons; the LAK dataset contains the description and content of scientific publications, information related the authors of the papers and the organization they are affiliated with; similar data types are contained in the L3S DBLP dataset though covering a broader research field.

[13] http://data.linkededucation.org/request/lak-conference/sparql.

[14] http://data.cnr.it/sparql-proxy/.

[15] http://data.open.ac.uk/query.

[16] http://dblp.l3s.de/d2r/sparql.

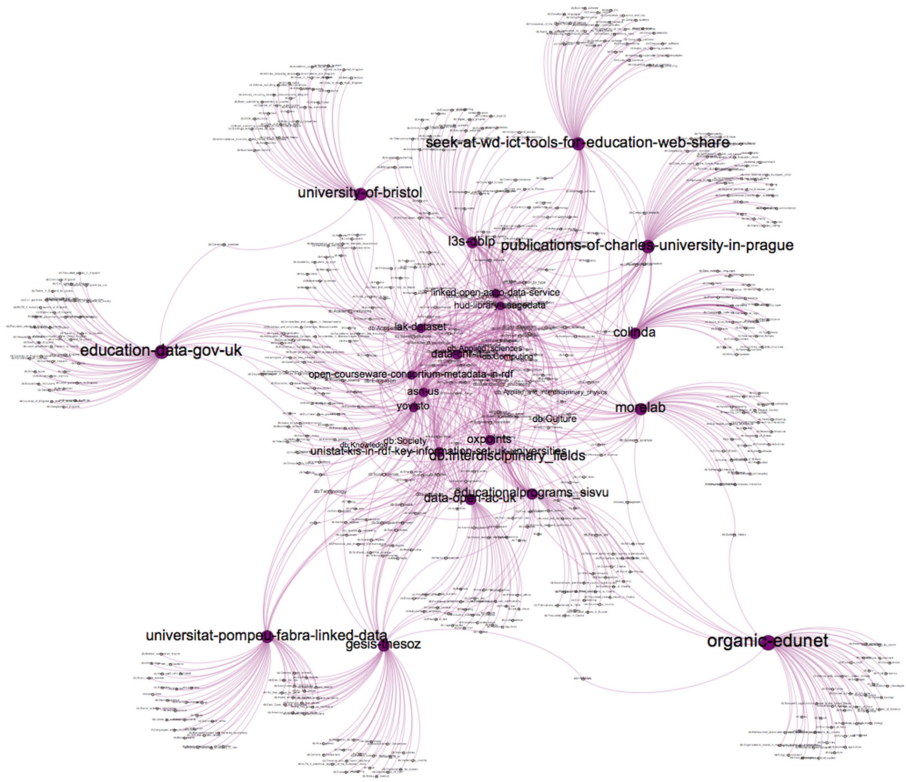

Fig. 1. Dataset and category graph (http://data-observatory.org/led-explorer/ch_fig_1.svg)

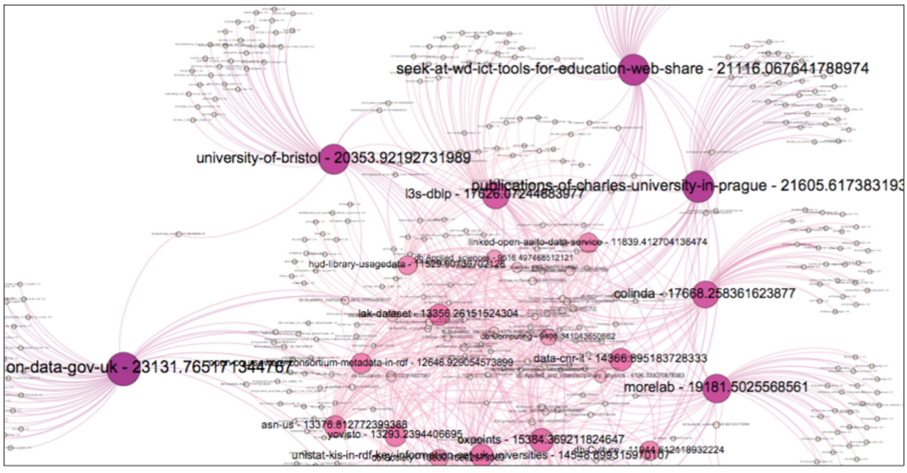

Fig. 2. Detailed view of the dataset and category graph

Finally, the part of The Open University (UK) dataset collects learning materials in different disciplines. Table 4 reports for these four datasets the associated categories with the highest relevance scores.

Table 4. Comparing top-k categories for four datasets of the Linked Educational cloud

k	Lak Dataset	Data.Cnr.it	data-open-ac-uk	L3S DBLP
1	Applied_sciences	Computing	Scientific_disciplines	Data_management
2	Educational_organizations	Interdisciplinary_fields	Interdisciplinary_fields	Information_retrieval
3	Academic_institutions	Data	Applied_sciences	Education
4	Educational_assessment_ and_evaluation	Information_technology	Science	Digital_media
5	Accreditation	Scientific_disciplines	Academic_disciplines	Book_websites
6	Education	Society	Knowledge	Archives
7	Applied_disciplines	Social_sciences	Natural_sciences	Digital_libraries
8	Computing	Data_management	Chemistry	Electronic_publishing
9	Academic_institutions	Academic_disciplines	Physical_sciences	Digital_library_projects
10	Digital_technology	Project_management	Education	Educational_projects
11	Academic_disciplines	Biology	Psychology	Computing_and_society
12	Computer_science	Land_Management	Branches_of_philosophy	Online_content_distribution
13	Data	Computer_data	Biology	Bibliographic_databases

It is important to highlight that even though these datasets have in common several categories, they contain different types of resources. Therefore, clustering mechanism based only on the categories shared by the dataset are not precise. For example, in the selected datasets, the *Biology* category are related with the data.cnr.it and the Open UK datasets even if these two datasets contain resources of entirely different types.

In Fig. 3 the effect of considering the resource type in the analysis of the relationships between dataset and categories is shown, with respect to the four datasets. Indeed, when the *foaf:Document* is considered only the data.cnr.it and lak-dataset are connected.

By considering the *foaf:Agent* the lak-dataset and l3 s_dblp shared categories, finally by considering the *aiiso:KnowledgeGrouping* only the data-open-ac-uk is connected to categories. This result is clearly related to the nature of the datasets analysed. The influence of resource type on dataset relationships is detailed in the next section.

3.3 Resource Type Coverage - the Dataset-Type-Graph

The type of the resources plays a key role to guide the exploration of the datasets and furthermore it is a strong indicator for the connectivity between datasets. Therefore, the influence of the resource types in the relationships between datasets has been investigated further. The Table 5 reports the ten resource types most frequently occurring across the 21 datasets under investigation.

Fig. 3. The effect of resource types on topic connections between datasets (a) all resource types; (b) foaf:Document; (c) foaf:Agent; (d) aiiso:KnowledgeGrouping (http://data-observatory.org/led-explorer/ch_fig_3{a,b,c,d}.svg)

The relationships between datasets and resource type have been analysed by means of the graph in Fig. 3 where nodes are resource types and datasets, an edge connects a dataset with a resource type if the dataset contains resources of that particular type.

In Fig. 4, four disconnected networks are represented. This is due to the fact that some datasets use specific vocabularies for their resources. Moreover, in the construction of this network the semantic relationships between resource types have not been taken into consideration. Consequently, types such as *foaf:Agent* and *foaf:Person* that are related by a *rdfs:subclass* relationship have been considered as distinct types. In order to improve the analysis of the relationships between datasets and resource types both existing mappings and new ones have been introduced. As existing mapping we consider the relationships that can be inferred from explicitly declared statements

in the vocabulary used in the datasets. Moreover, in the context of the LinkedUp project[17] a set of additional mappings has been introduced which link equivalent or overlapping types through standard OWL and RDF predicates, such as, *owl:equivalentClass* or *rdfs:subTypeOf*. A detailed description of the process that has led to the definition of these mappings is described in [2].

The following Fig. 5 reports two examples of mappings for the class *foaf:Agent* and *foaf:Document*.

Table 5. Most frequent resource types across educational Linked Datasets

Resource Type	#Datasets
http://xmlns.com/foaf/0.1/Document	5
http://xmlns.com/foaf/0.1/Person	5
http://www.w3.org/2004/02/skos/core#Concept	4
http://xmlns.com/foaf/0.1/Agent	3
http://purl.org/vocab/aiiso/schema#Institution	3
http://xmlns.com/foaf/0.1/Organization	3
http://rdfs.org/ns/void#Dataset	3
http://purl.org/vocab/aiiso/schema#Course	3
http://purl.org/vocab/aiiso/schema#Department	2
http://swrc.ontoware.org/ontology#InProceedings	2

The consideration of the mappings between resource types has made possible the aggregation of nodes representing resource types. In particular, equivalent resource types as well as resource types with subclass relationships have been grouped. The following Table 6 shows the most frequent resource type across the datasets after inference on mappings. In bold, we show the super-type, while the non-bold types indicate the most specific type association.

In Table 6, the resource types highlighted in bold represent the resource type together with all the resource type connected to it by considering the mapping. This table provides a clearer overview of the resources included in the datasets under investigation. Specifically, the most represented resource types (including also all of its subtypes, mapped types, inferred types) are related to *foaf:Document* (for instance, scientific and academic publications, educational resources), *foaf:Agent* (some of the datasets under investigation contain information about organizations, institutions and people) and *aiiso:KnowledgeGrouping*[18], since this class represents resources related to courses, learning modules, and so on. Type mappings across all involved datasets link "documents" of all sorts to the common *foaf:Document* class, "persons" and "organisations" to the common *foaf:Agent* class, and courses and modules to the *aiiso:KnowledgeGrouping* class.

Figure 6 represents the graph of dataset and resource types and their inferred types.

[17] http://linkedup-project.eu.

[18] http://purl.org/vocab/aiiso/schema#KnowledgeGrouping.

Fig. 4. Dataset and resource type graph without considering mapping (http://data-observatory.org/led-explorer/ch_fig_4.svg)

The density of this graph is lower than the one not including mappings, but the connectivity is higher. Subgraphs are stronger connected as in the previous network in which mappings are not considered. Indeed, in the network of Fig. 4, the graph density measured by gephi is lower (0.005) than in this case (0,011).

In Table 7, the datasets containing types linked to either or more of the three super-types are listed.

3.4 Type-Topic Correlation

As shown above, the resource type has a strong impact on the nature and semantics of the associated categories. While actual knowledge resources, such as OER, tend to be linked to explicit domains or disciplines, such as *Biology* or *Computer Science*, the range of categories for persons and organisations is of entirely different nature. While topics/categories are always linked to particular resources and their types, the joint analysis of both types and topics is of crucial importance to enable a better understanding of educational Linked Data. Considering the resource types associated

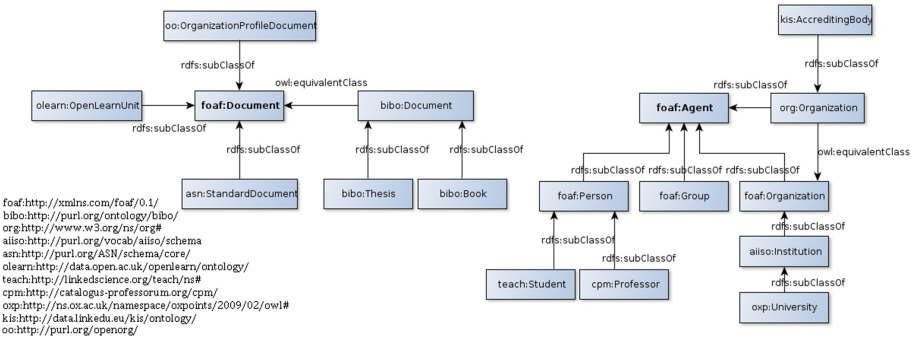

Fig. 5. Mappings between resource types related to foaf:Agent and foaf:Document

Table 6. Most frequent resource types according to their representation in the datasets (mapping considered)

Resource Type	# Datasets
foaf:Agent	**14**
foaf:Person	5
foaf:Organization	3
aiiso:Institution	3
foaf:Agent	3
aiiso:Department	2
foaf:Document	**12**
foaf:Document	5
bibo:Article	2
bibo:Book	2
bibo:Document	2
swrc:Document	2
swrc:InProceedings	2
aiiso:KnowledgeGrouping	**7**
aiiso:Course	3
aiiso:Module	2
courseware:Course	2
skos:Concept	**6**
skos:Concept	4
geo:SpatialThing	**4**
c4 dm:Event	**3**
void:Dataset	**3**

with each topic in the dataset topic profile graph, it has been possible to create a network in which the resource types have been connected with the categories they are related with.

In the graph of Fig. 7, seven groups of nodes are clearly identified. The following table reports the most representative categories related with the most connected

Fig. 6. Datasets and resource type graph considering inferred types (http://data-observatory.org/led-explorer/ch_fig_6.svg)

Table 7. Datasets for the most frequent resource types (considering type mappings)

foaf:Agent	foaf:Document	aiiso:KnowledgeGrouping
morelab	morelab	universitat-pompeu-fabra-linked-data
colinda	organic-edunet	asn-us
publications-of-charles-university-in-prague	asn-us	unistat-kis-in-rdf-key-information-set-uk-universities
lak-dataset	open-courseware-consortium-metadata-in-rdf	oxpoints
university-of-bristol	universitat-pompeu-fabra-linked-data	linked-open-aalto-data-service
l3 s-dblp	university-of-bristol	data-open-ac-uk
oxpoints	lak-dataset	educationalprograms_sisvu
education-data-gov-uk	yovisto	
educationalprograms_sisvu	l3 s-dblp	
linked-open-aalto-data-service	oxpoints	
unistat-kis-in-rdf-key-information-set-uk-universities	linked-open-aalto-data-service	
	hud-library-usagedata	

resource types in the LinkedUp catalog. In order to enable a better distinction, we particularly consider the most frequent resource types including the resource types associated to them by means of mappings.

Table 8 provides evidence that the resource types related to person and organization (*foaf:Agent*) are more connected to physical places and locations, while resource types related to actual documents (*foaf:Document*) or courses (*aiiso:KnowledgeGrouping*) are representing actual domains and disciplines. For the latter, we observe a strong bias towards topics relating to Computer Science and the Life Sciences. Regarding the *foaf:*

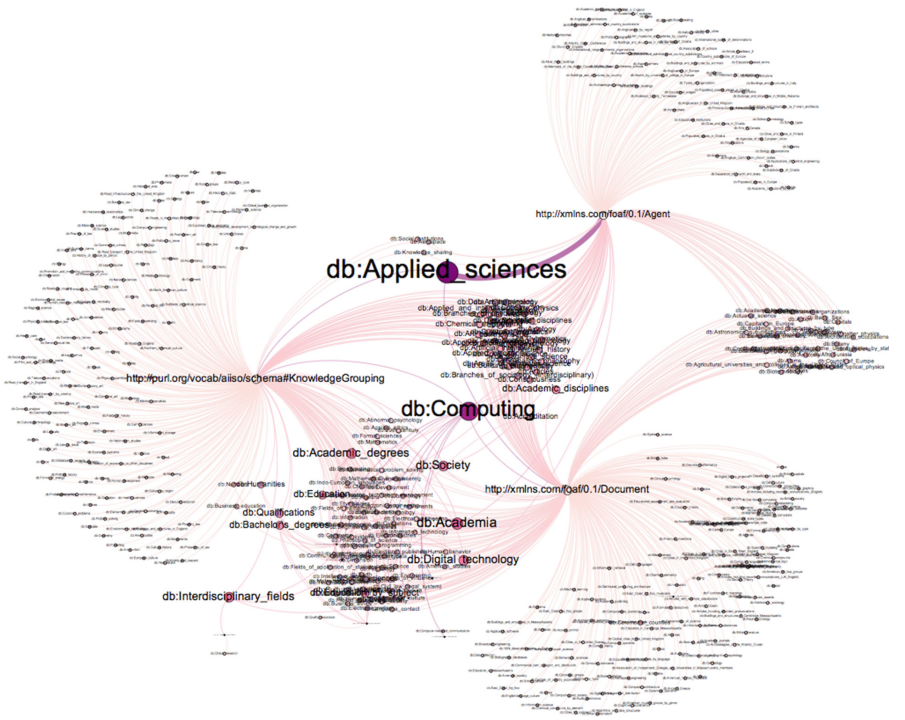

Fig. 7. Resource types and categories (http://data-observatory.org/led-explorer/ch_fig_7.svg)

Organization it is possible to highlight a bias towards the categories related to European area such as: *Cities_in_Europe*, *Baltic_Sea*, *Capitals_in_Europe*.

This observation correlates with the general intuition that such topics are also stronger represented in the Linked Open Data cloud in general and might lead to additional research into how to resolve such a topic bias in the future.

This observation can be repeated when analysing a single dataset. For instance in the case of the University of Bristol[19] dataset, that contains resources related to documents and agents, the category distribution among the most frequent resource types/categories is shown in Table 9. The top ten categories according to the score calculated in the profile are reported in the table.

In general, when the categories related to a dataset are analysed, the distribution of the categories is usually heterogeneous if the resource type is not considered. When the resource type is taken into account, so that the types where they originate from (document, person, organization) are identified, the category subgraph pertaining to a specific resource type usually can is more coherent, distinct and connected. This observation can be exploited, for instance, when aiming to automatically map resource types and identifying type relationships, where the associated category overlap or similarity can serve as indicator.

[19] http://resrev.ilrt.bris.ac.uk/data-server-workshop/sparql.

Table 8. Most frequent categories for most frequent resource types

foaf:Document		foaf:Agent		aiiso:KnowledgeGrouping	
Applied_sciences	1164	Applied_sciences	1522	Digital_technology	1393
Biology	680	Academic_institutions	533	Computing	1262
Academic_disciplines	656	Academic_disciplines	823	Society	1011
Branches_of_philosophy	624	Educational_organizations	533	Interdisciplinary_fields	793
Chemistry	604	Types_of_organization	523	Education_by_subject	789
Areas_of_computer_ science	593	School_types	520	Academia	717
Education	591	Schools	520	Academic_disciplines	688
Artificial_intelligence	581	Organizations	520	Education	653
Computing	548	Educational_institutions	520	Applied_sciences	648
Branches_of_psychology	548	Educational_buildings	516	Qualifications	591

Table 9. Distribution of top-k categories for the University of Bristol dataset

k	All resource types	foaf:Document	foaf:Agent	
			foaf:Person	foaf:Organization
1	db:Academic_disciplines	db:Academic_disciplines	db:Buildings_and_structures_ by_type	db:Academic_disciplines
2	db:Buildings_and_structures_ by_type	db:Biology	db:Chemistry	db:Buildings_and_structures_ by_type
3	db:Biology	db:Applied_sciences	db:Cities_in_Europe	db:Biology
4	db:Applied_sciences	db:Chemistry	db:Capitals_of_country_ subdivisions	db:Applied_sciences
5	db:Academic_publishing	db:Articles_including_recorded_ pronunciations_(English)	db:Capitals	db:Chemistry
6	db:Academic_publishing_ companies	db:Articles_including_recorded_ pronunciations	db:Arts_in_the_United_Kingdom	db:Business_organizations
7	db:Chemistry	db:Articles_including_recorded_ pronunciations_(UK_English)	db:Capitals_in_Europe	db:Building_engineering
8	db:Cities_in_Europe	db:Academia	db:Cities_in_the_United_States	db:Branches_of_philosophy
9	db:Academia	db:Academic_journals	db:Academic_institutions	db:Anatomy
10	db:Cities_in_South_West_ England	db:Book_publishing_companies_ by_country	db:British_capitals	db:Architecture

4 An Interactive Explorer for Educational Linked Data

In this section, we present the *dataset profile explorer* developed with the aim of supporting an effective re-use of the resources in the educational LD cloud. In particular, the *explorer* makes explicit the topics covered by the datasets even in relation to the types of resources, mainly focused on the domain of educational related datasets. As stated in previous Sect. 3, topic coverage and the type of the resources in this domain assume a key role in supporting the search for content suitable for a specific learning course. The *explorer* allows users to navigate topic profiles associated with datasets with respect to the type of the resource in the dataset.

As shown in Fig. 8, the *explorer*[20] is composed of three panels: the panel at the centre of the screen shows the network of datasets and categories; the panel on the left shows general and detailed descriptions about datasets and categories; the *selection*

[20] http://data-observatory.org/led-explorer/.

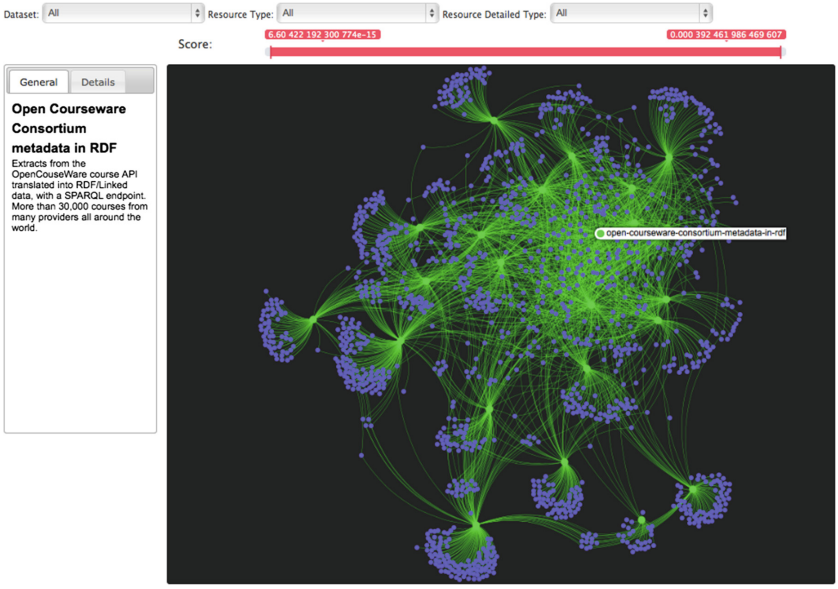

Fig. 8. A screenshot of the demo

panel, placed at the top of the screen, allows users to apply specific filters on the network. In the central panel, green nodes represent datasets while blue nodes represent categories. An edge connects a dataset to a category if the category belongs to the dataset topic profile. In order to draw the network, the *sigmajs*[21] library has been used and the nodes of the network have been displayed using the ForceAtlas2 layout. By clicking on a node (dataset or category), general and detailed descriptions are shown on the left panel. In the case of a dataset, the general description reports the description of the dataset retrieved from the *datahub.io* repository[22]. In the detailed description, the list of the top ten categories (and the related score) associated to the dataset is reported. In the case of a category, the description panel reports the list of datasets which have that category in their profile. The datasets including the category in their top ten list are highlighted in bold.

The selection panel allows users to filter the results according to: dataset, resource type, and resource *sub-type*. The list of datasets is composed by the datasets of the LinkedUp catalog. Regarding the resource type, the *explorer* is focused on three main classes: *foaf:Document*, *foaf:Agent* and *aiiso:KnowledegeGrouping*. As reported in Subsect. 3.4, these three classes are the most represented classes in the datasets, and *foaf:Document* is related to learning material such as: research papers, books, and so on; the *foaf:Agent* resource type has been included to take into account elements such as persons and organizations. The *aiiso:KnowledegeGrouping* is a type representing

[21] http://sigmajs.org.

[22] http://datahub.io.

resources related to courses and modules. This initial set of resource types can be easily enlarged by means of configuration settings. The resource sub-type has been included with the aim of refining the results already filtered by resource type. Another filter is related to the score of the relationships between datasets and categories. A slider bar allows users to filter results based on a specific range of the scores, calculated by the linked dataset profiling pipeline [3]. The filters on datasets, resource types and resource sub-types can be combined and, as a result, only the portion of the network consistent with the filter selection is highlighted. Even though the *explorer* has been tested with an initial group of datasets of the LinkedUp Catalog, it can be configured in order to extend the number of datasets covered. Moreover, the *explorer* can cover also datasets on different fields provided that the dataset topic profile is available, thus extending the application of the proposed approach to several fields.

5 Conclusion

In this work, we have provided an analysis of the coverage of educational Linked Data on the Web and an investigation of the inherent correlations between types, topics and datasets. Only the joint consideration of types and topics allows the non-ambiguous interpretation of topic annotations of datasets. Key findings of our study include:

F1. Educational datasets can best be characterised (profiled) by a combined representation of resource types and categories as part of dataset profiles.
F2. The nature of categories differs significantly depending on the resource types they are associated with. In other words, the distinct subgraphs of the DBpedia category graph characterise resources of very distinct types.
F3. Educational and presumable cross-domain resource types can be characterised by their inherent topic distribution.
F4. Educational resources, i.e. instances which represent some form of educational documents, currently are not equally spread across all disciplines. A topic bias exists towards fields in the area of Computer Science and the Life Sciences.

Our analysis uncovers an inherent topic bias of educational resources represented in datasets, usually focused on disciplines related to *Computer Science* and *Life Sciences*, where for instance, social sciences appear to be underrepresented. While this bias emerged for specific resources, i.e. instances by types which can be summarised as some notion of *document*, including dedicated OER, scholarly papers or audiovisual material, a similar bias was not detected for other types such as organisations or persons. In such cases, a deeper analysis, for instance of the origins and characteristics of represented entities, taking into account background knowledge such as geodata, seems better suited to detect some form of demographic or geographic scope or bias. As shown above, the nature of categories associated with resources of different types differs significantly depending on the respective resource type. For instance, while actual document-related types usually are related to topics which indicate some form of subject or domain (such as "*Biology*"), resources representing some notion of agent, such as organisation or person, usually are characterised through some broader categorisations, such as "*Academic_institutions*" or "*People_from_Athens*".

While this suggests that the subgraphs of the DBpedia category graph tied to specific resources fundamentally differs, we argue that category distributions of resource types might provide a useful means for mapping and aligning types (F3). The intuition is that similar categories are likely to be tied to instances of similar resource types. While type mappings in the educational Linked Data landscape as well as the LinkedUp data catalog currently are mostly created manually by experts, as part of future work we are investigating possibilities to exploit this observation as part of automated type and schema alignment methods.

Acknowledgments. This work has been partially supported by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No 317620– LinkedUp project (http://linkedup-project.eu/). The authors would like to acknowledge networking support by the COST Action IC1302 (KEYSTONE).

References

1. Fetahu, B., Dietze, S., Nunes, B.P., Taibi, D., Casanova, M.A.: Generating structured profiles of linked data graphs. In: 12th International Semantic Web Conference (ISWC2013), Sydney, Australia (2013)
2. D'Aquin, M., Adamou, A., Dietze, S.: Assessing the educational linked data landscape. In: ACM Web Science 2013 (WebSci 2013), Paris, France, May 2013
3. Fetahu, B., Dietze, S., Pereira Nunes, B., Antonio Casanova, M., Taibi, D., Nejdl, W.: A scalable approach for efficiently generating structured dataset topic profiles. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 519–534. Springer, Heidelberg (2014)
4. Dietze, S., Yu, H.Q., Giordano, D., Kaldoudi, E., Dovrolis N., Taibi, D.: Linked education: Interlinking educational resources and the web of data. In: ACM Symposium on Applied Computing (SAC-2012), Special Track on Semantic Web and Applications (2012)
5. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web (LDOW2008). In: Proceedings of the 17th International Conference on World Wide Web (WWW 2008), April 21-25, 2008, Beijing, China (2008)
6. Bizer, C., Heath, T., Bernes-Lee, T.: Linked data - the story so far. Special Issue on Linked data. International Journal on Semantic Web and Information Systems (IJSWIS) (2009)
7. Heath, T., Bizer, C.: Linked data: evolving the web into a global data space (1st edition). In: Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1–136. Morgan & Claypoo (2011)
8. de Santiago, R., Raabe, A.L.A.: Architecture for learning objects sharing among learning institutions-LOP2P. IEEE Trans. Learn. Technol. **3**, 91–95 (2010)
9. Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H.Q., Giordano, D., Marenzi, I., Pereira, N. B.: Interlinking educational resources and the web of data: a survey of challenges and approaches. Emerald Program: Electron. Libr. Inf. Syst. **47**(1), 60–91 (2013). doi:10.1108/00330331211296312
10. Taibi, D., Dietze, S.: Fostering analytics on learning analytics research: The LAK dataset. In: CEUR WS Proceedings vol. 974, Proceedings of the LAK Data Challenge, held at LAK2013 – 3rd International Conference on Learning Analytics and Knowledge (Leuven, BE, April 2013) (2013)

11. Taibi, D., Dietze, S., Fetahu, B., Fulantelli, G.: Exploring type-specific topic profiles of datasets: A demo for educational linked data. In: Poster & System Demonstration Proceedings of 13th International Semantic Web Conference (ISWC 2014), Riva Del Garda, Italy, October 2014
12. Mitsopoulou, E., Taibi, D., Giordano, D., Dietze, S., Yu, H.Q., Bamidis, P., Bratsas, C., Woodham, L.: Connecting medical educational resources to the linked data cloud: The mEducator RDF schema, store and API, in linked learning 2011. In: Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age, CEUR-WS, vol. 717 (2011)
13. UNESCO. Forum on the impact of Open Courseware for higher education in developing countries. Final report. Paris: UNESCO (2002)
14. Geser, G.: Open Educational Practices and Resources. OLCOS Roadmap (2012)
15. Hylén, J.: Open educational resources: Opportunities and challenges. In: Proceedings of Open Education 2006: Community, Culture and Content, pp. 49–63 (2006)
16. Atkins, D.E., Brown, J.S., Hammond, A.L.: A review of the Open educational Re- sources (OER) movement: achievement, challenges and new opportunity. Report to the William and Flora Hewlett Foundation (2007)
17. OECD: Giving Knowledge for free: the Emergence of Open Educational Resources. OECD, Paris (2007)
18. Taibi, D., Fulantelli, G., Dietze, S., Fetahu, B.: Evaluating relevance of educational resources of social and semantic web. In: Hernández-Leo, D., Ley, T., Klamma, R., Harrer, A. (eds.) EC-TEL 2013. LNCS, vol. 8095, pp. 637–638. Springer, Heidelberg (2013)

http://www.springer.com/978-3-319-30492-2

Open Data for Education
Linked, Shared, and Reusable Data for Teaching and Learning
Mouromtsev, D.; d'Aquin, M. (Eds.)
2016, VII, 189 p. 76 illus., Softcover
ISBN: 978-3-319-30492-2