future internet

MDPI

Article

Digital Libraries: The Challenge of Integrating Instagram with a Taxonomy for Content Management

Simona Ibba * and Filippo Eros Pani

Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, 09125 Cagliari, Italy; filippo.pani@diee.unica.it

* Correspondence: simona.ibba@diee.unica.it; Tel.: +39-070-675-5774

Academic Editor: Wolf-Tilo Balke

Received: 29 February 2016; Accepted: 25 April 2016; Published: 10 May 2016

Abstract: Interoperability and social implication are two current challenges in the digital library (DL) context. To resolve the problem of interoperability, our work aims to find a relationship between the main metadata schemas. In particular, we want to formalize knowledge through the creation of a metadata taxonomy built with the analysis and the integration of existing schemas associated with DLs. We developed a method to integrate and combine Instagram metadata and hashtags. The final result is a taxonomy, which provides innovative metadata with respect to the classification of resources, as images of Instagram and the user-generated content, that play a primary role in the context of modern DLs. The possibility of Instagram to localize the photos inserted by users allows us to interpret the most relevant and interesting informative content for a specific user type and in a specific location and to improve access, visibility and searching of library content.

Keywords: taxonomy; metadata standards; digital libraries; Instagram; hashtag; user-generated content

1. Introduction

What are digital libraries (DLs)? According to [1], the term "digital library" is used in several distinct senses and is used to describe a variety of entities and concepts. The main definition of DLs is made up of two fundamental aspects [2]:

1. DLs are correlated technical potentialities to create, search and use information. They are an extension of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds, static or dynamic images) and exist in distributed networks. The content incorporates some metadata that describe various aspects and others that consist of links to data or metadata, whether internal or external to the DL. Metadata play a key role in the organization and management of digital resources, especially when the amount of available information is high and must be indexed and catalogued for easy search and retrieval. Metadata must make the resource accessible by adding tags to the content according to a consistent pattern [3];
2. DLs are built and organized by users, and their functional capabilities support the information uses of users. They are an extension of information institutions where resources are collected, organized and accessed.

The two above definitions explain what the current challenges are in the context of DLs. Therefore, the most important aspects that need to be carefully handled in DLs are [4]: interoperability, document representation, intellectual property, usability, social and economic implications, supporting network infrastructure and scalability. In this paper, we want to focus on the first two aspects of interoperability and social implication. DLs allow and facilitate the exchange of information and knowledge among people. In this sense, DLs also have a social purpose of bringing together people around educational themes.

Knowledge can be seen, from an operational point of view, as a valid certainty which improves the abilities of a man to undertake efficient actions. The management of this knowledge means defining a process of applying a systematic approach to structure knowledge and making it available for sharing and reuse [5].

The use of social networks is a very important method of sharing knowledge in the context of DLs. As Houghton-John [6] defines it, DLs "simply means making a library's space (virtual and physical) more interactive, collaborative, and driven by community needs".

The definition can be interpreted in light of the relevance of social participation among a DL's community. That very important aspect includes a lot of instruments such as social networking, tagging, blogging, social bookmarking, podcasting and so on. Among these tools, we chose to use the mobile application Instagram that allows users to insert hashtags related to the images. Each picture published on Instagram also has some metadata associated with it.

In the DL context, there are heterogeneous resources (images, music, audio, databases, ebooks, audiobooks, websites, pages or social network accounts), and some of these are unstructured or described with different metadata schemas both in the analysis and in the attributes assigned to resources. The integration of resources is a complex activity due to the quantity of existing metadata schemas. Our work aims to find a relationship between the main metadata schemas by comparing them. In particular, our aim is to formalize knowledge through the creation of a metadata taxonomy built through the analysis and the integration of existing metadata schemas associated with DLs and through the interpretation and the combining of Instagram metadata and Instagram hashtags.

We discussed the following questions:

1. Is interoperability between metadata and the user-generated content (UGC) with Instagram possible?
2. Does the use of Instagram improve resource management of DLs?

The final result of our work is a taxonomy, which provides innovative metadata with respect to the classification of resources, as images of Instagram and the UGC. The application must provide innovative features related to semantic search; in particular, it will be capable of properly managing the UGC related to library resources; it also involves the creation of a social network.

The paper is structured as follows: Section 2 identifies the advantages of social media application in a DL context. Section 3 presents metadata standards for DL, Section 4 explains the Instagram metadata, whereas Section 5 describes UGC for DLs. Section 5 also presents an overview of the state-of-the-art, whereas Section 6 describes our approach for multimedia content management in the context of DLs. Sections 7–9 explain the method of constructing the taxonomy. Section 10 describes the structure of the resulting taxonomy, whereas in Section 11, the discussion about the use and the future evolution of the work is presented. The conclusion is provided in Section 12.

2. Social Media Application in a Digital Library Context

The efforts to improve access to DLs will focus on two key routes: interoperability and the importance of content posted by users. An information gap may exist between the formal information entered by the operators of a DL and the users' perspectives. As with access to digital museums, according to [7], the people "search for meaning: not just records". According to [8], while many collections are appreciated for the quality of their objects and preservation techniques, they often remain inaccessible to their users. One of the ways digital librarians can acquire a broader awareness of their collections is through social networking. Social media then allow the DLs to make information more accessible: the UGC and the tagging systems associated with digital resources are valid methods to bridge the information gap and to make the approach to knowledge more collaborative and interoperable.

Within a DL's strategic plan, we want to highlight two determinative aspects that provide digital librarians with direct strategies for successfully integrating social media: understanding of the content posted by users and integration with a formal knowledge management. Digital librarians need to find users where they are located. The first step is to find out where the conversations are happening, who are the stakeholders and what is the satisfaction of users with regard to a particular resource. It is important, for example, to find out what are the most interesting topics for users in a specific geographical region. The DL can even be useful to identify what are the most central accounts about a specific topic.

Thus, the user becomes a contributor: when he tags some information, it benefits the community. Social tagging is key to the process of collaboration of users: it is a personal free tagging of information and resources for one's own retrieval. The tagging is done in a shared social environment. The result of this tagging is shared in the community and it produces knowledge [9]. This system lets users organize the information in a way that they can easily retrieve the information. People have surprising ways of tagging information: this could lead to cross-links between information that would never have come to light without the use of system of tagging.

In our work, among the social media that integrate a tagging system, we chose to use Instagram. We use Instagram because it has a large volume of publicly accessible data—no need to login, no need to be friends to see someone's photos (except for private accounts)—and provides valuable information to locate users. This element is very important to create location-based relationships between the resources of DLs. Instagram is described in the next subsection.

Instagram

Instagram is a mobile application that started in 2010, and, in 2015, it surpassed the milestone of 300 million registered users [10]. Instagram has different features from the other tools: it is an exclusively mobile application, it does not allow the posting of direct links to sites. Instagram allows the upload and sharing of photos and videos through the use of a mobile device. It offers its users a unique way to post their pictures, and allows immediate editing through 22 filters. It also allows users to add captions, hashtags using the # symbol to describe photos or videos, or to mention other users using the @ symbol (the @ symbol creates an actual link among the accounts). Instagram also allows users to follow posts from all selected profiles, and to have their own followers. Each follower must be specifically approved by the profile owner. Every user can set their own privacy preferences, and can make the pictures visible to everyone or only to followers. Pictures in profiles are shown in chronological order, starting from the latest. For each picture, it is possible to enter likes or comments. Hashtags and user mentions can be entered inside comments. Moreover, according to research by [11], Instagram photos can be categorized into eight types based on their content: activities, self-portraits, captioned photos, friends, food, fashion, gadgets, and pets. Instagram is even an aware-location application [12].

The Instagram Application Programming Interface (API) allows queries around user-specified tags, providing extensive information about relevant images and videos for searches around particular hashtags or keywords [13]. The information provided allows for the analysis of collected data to incorporate several different dimensions; for example, the information about the tagged images returned through the Instagram API allows us to examine patterns of use around publishing activity (time of day, day of the week), types of content (image or video), and locations specified around these particular terms. The Instagram API gives researchers access to metadata containing a lot of interesting information for this work.

3. Metadata Standards for Digital Libraries

The traditional methods of cataloguing of bibliographic formats allow for accurate access to resources and their location, but the quantity, variety and increase in the amount of digital content produces new instruments for the management and description of these resources. The use of metadata is a solution to retrieve digital objects in a precise way through a single point of access. Metadata describe the structure, features, content, conditional use and ways for managing the library resources. Thus, metadata are fundamental instruments for the retrieval of information and to guarantee complete interoperability [14]. The metadata used in the DL context must provide features to:

- identify and find the resources (descriptive metadata);
- manage resources and guarantee their acquisition, in accordance to the related rights and licenses (metadata management);
- relate a resource's components to make the information accessible (structural metadata).
- Metadata standards are either application-specific or generic. Often, metadata standards provide information only for a particular type of multimedia object, or for a restricted set of multimedia objects, only for specific image and audio files. Below, we provide a brief introduction to the most relevant standards.

3.1. Dublin Core Standard

The Dublin Core (DC) is a standard that originates from the need to promote the cohesion of different digital resources, through their identification with a limited number of attributes. It is a vocabulary suitable to the basic communication that still allows the discovery of resources with an appropriate level of accuracy. The core of the standard is composed by 15 elements, and is part of a larger set of metadata vocabularies and technical specifications maintained by the Dublin Core Metadata Initiative (DCMI) schema. The simple DC (without 'qualifiers'), is used for the exchange of metadata according to the OAI-MHP (Open Archive Initiative Protocol for Metadata Harvesting) [15]. The elements defined in the schema are: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type. The need to express certain values more accurately has led to the definition of the qualifiers. The full set of vocabularies (DCMI Metadata Terms) also includes a set of resource classes including the DCMI Type Vocabulary, vocabulary encoding schemas, and syntax encoding schemas. The schema can be extended by defining additional elements appropriately identified by a prefix that indicates the schema to which it belongs. Through an application profile, an additional metadata can be inserted, typical of the context and not covered by the basic schema, and also technical and management metadata, useful for the management of resources. The DC standard is therefore a model of organization that aims at ensuring effective collaboration through the sharing of common principles. Thus, it helps to achieve global access to information while preserving the specificity of the resources. The main features of DC are the ease of use, the semantic interoperability and the flexibility, as it allows you to integrate and develop the data structure with different semantic meanings.

3.2. XMP Standard

The Adobe Extensible Metadata Platform (XMP) is a standard for processing and storing information relating to the content of a file [16]. XMP standardizes the definition, creation and processing of extensible metadata. Serialized XMP can be embedded into file formats: embedding metadata allows for avoiding several issues that occur when metadata are stored separately.

XMP is used in PDF and photo editing applications. This standard encapsulates metadata inside the file, using Resource Description Framework (RDF), a basic tool proposed by World Wide Web Consortium (W3C) to encode, exchange and reuse the structured metadata.

3.3. Exif Standard

Exchangeable image file format (Exif) is a standard created by Japan Electronics and Information Technology Industries to specify the formats of digital systems handling image and sound files [17]. Exif is supported by the main producers of digital cameras and it gives users the capability of embedding pictures with information that can be used by various devices to improve processing.

Exif offers a set of specific tags: these cover a wide spectrum including: time and date information, memorizing the current date and time; camera settings, containing static information about the camera's model and producer, information about the orientation, aperture, shutter click speed, focal length, white balancing and International Organization for Stardardization (ISO) speed information for every image; information about the shutter click's location coming from a GPS receiver connected to the camera; and information and descriptions about the copyrights.

4. Instagram Metadata

The metadata of Instagram provides relevant and accurate information on user name, date and time of image creation, the location where the picture was taken, the caption entered by the author, comments, hashtags associated with the image, number of likes and names of the users that gave their like. In addiction to the metadata associated with images, metadata associated with each user that has posted a picture can be extracted. This metadata allows us to find the number of followers and following, email address, number of posts and a brief biography. The high quantity of available data allows us perform quantitative and qualitative analysis, to verify the stream of content over time and find the most interesting topics to users. The information gathered from it makes it also possible to map data according to their geographic location, and to analyze the geolocalization of users in relation to specific tags of interest. The analysis on tags allows us find users' consumption models, the type of posted content, and the specific locations where the same content is posted, together with time data. These elements are very important for integration with DLs.

Instagram Hashtags as Image Annotation Metadata

In addition to the metadata that the Instagram application provides by default [18], it demonstrates that Instagram Hashtags can be interpreted as Image Annotation Metadata.

In their work, they prove that tags accompanying photos in social media and especially the Instagram hashtags, provide a form of image annotation and that Instagram hashtags, and especially those provided by the photo owner/creator, express more accurately the content of a photo compared to the tags assigned to a photo during explicit image annotation processes like crowdsourcing. It has also been found that both the image content and the context in which an image resides affect interpretability, but by measuring whether the other people would choose the same hashtags with the image creator/owner they found that, in 55% of the chosen hashtags, participants and owners agree that the suggested hashtags can describe the visual content of an image. These results lead us to integrate the Instagram hashtag within our taxonomy.

5. User-Generated Content for Digital Libraries

According to [19], DLs and UGC are two expressions closely connected. This work proves that the social tagging is an important method for web users to add keywords to online objects of a library 2.0 and to improve the access to digitalized objects and to index resources.

In [20], an approach is shown instead to integrate text and web mining with social tagging systems that altogether provide semantic search as a service of DLs.

Ulwazi Programme is a community-generated DL of local content [21]. This project employs crowdsourcing and Web 2.0 technologies to enable local communities to contribute to a DL. The Taiwan Digital Library is a history library with the reading annotation tool for knowledge sharing [22]. This project explains how to implement such DL systems and how the UGC benefits the growth of digital archives. In [23], a DL of movie review documents was developed that supports sentiment-based browsing and searching by UGC. Using the system, you can browse and search movies by analysis of users' sentiment.

6. Related Work

A relevant work is [24] that has analyzed the requirements for importing documents and metadata into DLs and described a new extensible architecture that satisfies these requirements. The proposed structure converts heterogeneous document and metadata formats, organized in arbitrary ways on the file system, into a uniform XML-compliant file structure. This simplifies the construction of the indexes, browsing structures, and associated files that form the basis of the runtime DL system. In accordance with [25], the ideal solution to metadata interoperability difficulties would be the adoption, strict adherence to, and consistent implementation of a single standard by all DL. Such an approach has been pursued by some libraries in the past as exemplified by the adoption of the Dewey Decimal Classification system, the Anglo-American Cataloguing rules (AACR2), and the Machine Readable Cataloguing (MARC), but such efforts have had their big problems. Furthermore, the existence of several metadata standards, coupled with the proliferation of several "in-house" schemas has exacerbated the situation. Under such circumstances, achieving metadata interoperability, with the adoption of a single standard, becomes a daunting task [26]. It is difficult to integrate different metadata standards because of the overlapping in functionality and their semantic ambiguity. Different standards are often not designed for a combined use. Many solutions have been proposed to provide a formal classification that could take into account the relationships between different multimedia metadata [27]. Ontologies based on the Moving Picture Experts Group-7 (MPEG-7) [28,29] standard, like the one proposed by [30], the one proposed by [31], and the MPEG-7 Upper Multimedia Description Schemes (MDS) [32] developed within the Harmony Project, all represented in Ontology Web Language (OWL), are not suitable for an immediate use in the Italian DL scenario, both for the higher emphasis placed on audio and video content than on other multimedia objects, and for the interoperability issues connected with the exploitation of the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). The Multimedia Metadata Ontology (M3O) [33] is another solution to metadata standard integration issues. M3O is a modelling framework that targets the multimedia metadata standard integration issues by abstracting from existing standards. The alignment method used by the creators of M3O does not use machine learning approaches; instead, M3O makes use of a pure manual alignment, in order to ensure a high quality of standard integration. Another relevant ontology to be considered is the Media Resource Ontology. Created by the W3C Media Annotation Working Group, it is an ontology based on the mapping definition of many different multimedia metadata standards, including Exif 2.2, MPEG-7, Metadata Encoding and Transmission Standard (METS) [34], National Information Standards Organization (NISO) [35] and XMP. It is mainly web-oriented, and, being structured following other standards, does not analyze the specific elements of the context from which the general concepts of taxonomy can be obtained. A remarkable effort to obtain a software-independent, and also hardware-independent, formalization, is the MAG (Metadati Amministrativi e Gestionali) standard [36]. The MAG schema is an application profile that interacts with other standards: DC and NISO. It is not like the modelling framework M3O, because the MAG schema defines a metadata taxonomy, so it is not as complex as an ontology, and it can achieve a higher degree of independence, both from application context, and from software and hardware. MAG metadata are specified through the XML format, in order to be compliant with the OAI-PMH standard. As an extensible standard, MAG could be a good starting point for the construction of a metadata taxonomy. Despite MAG being much easier to use than modeling frameworks like M3O, the approach used to build it lacks a phase analysis of existing DLs, on which the approach proposed in this paper is based. Another application profile, similar to MAG, is the PICO (Portale Italiano della Cultura Online) AP (Application Profile) [27]. PICO AP, like MAG, is oriented to the exploitation of OAI-PMH, and its target is to ensure metadata harvesting functionalities also in the presence of different schemas, in addition to reaching a future-proof structure and supporting interoperability. Like MAG, the PICO AP is an XML metadata schema which makes use of international standards. It is a DC application profile. PICO AP has been recently extended, via a specific encoding schema, to support the encoding of metadata provided by ICCD (Istituto Centrale per il Catalogo e la Documentazione—Central Institute for Cataloguing and Documentation) cards.

A recent work about taxonomies is [37] where the role of users is very important because they can declare different kinds of relationships among information objects of the library content. The resulting taxonomy is represented as a set of controlled semantic vocabularies of linkage classes.

In [38], a taxonomy of DL services was accomplished to improve issues of reusability, extensibility and composability. In this project, some applications of taxonomy were described to propose a modeling language for DLs and the specification of quality metrics to evaluate DLs.

Another significant work is [39] that examines the linguistic structure of folksonomy tags collected over a thirty-day period from the daily tag logs of Del.icio.us, Furl, and Technorati. The tags were evaluated compared to the NISO guidelines for the construction of controlled vocabularies. The results indicate that the tags correspond closely to the NISO guidelines pertaining to types of concepts expressed, the predominance of single terms and nouns, and the use of recognized spelling. The work showed that the folksonomies can serve as a powerful, flexible tool for increasing the user-friendliness and interactivity of public library catalogs, and also may be useful for encouraging other activities, such as informal online communities of readers and user-driven readers' advisory services.

The discussion about the use of folksonomies and other methods of classification is very interesting. Several works were developed to find interoperability processes between these kinds of formalization. Peterson asserts that folksonomies are based on relativism [40]. Moreover taxonomies consistently provides better results to users. Lee and Neal studied the way in which people give a specific tag to a photo: tags commonly are associated with objects in the photograph and events taking place in the photograph [41]. An interesting contribution is that proposed by [42]. He asserts that folksonomies are not in conflict with the taxonomies but are supplementary to them. Upper Tag Ontology (UTO) instead is a useful method to facilitate modeling of tagging data and the integration of data from different bookmarking sites (Delicious, Flickr and YouTube) and to study the interoperability of tagging ontologies [43].

7. Developed Approach

We dealt with the problem of interoperability of metadata standards and of tags inserted by users on Instagram with the adoption of a hierarchy of classes, that is, a taxonomy [44].

We started with the definition of the most general concepts and then we added more specialized concepts. The general contents are placed at a higher level, while concrete details are placed at a lower level. Starting from a formalization of the reference knowledge (taxonomy, metadata schema), we then classified the information found in the reference domain.

In the second step, we defined the most specific classes, (the leaves of the hierarchy) that were also obtained by grouping the hashtags inserted by users, with subsequent grouping of these classes into more general concepts.

In the end, we compared the concepts obtained with the two previous steps and we established the final structure of the taxonomy.

This is analyzed, using the available information, in order to define a reference terminology to describe the data.

Given that the main purpose of our work is to analyze the objects of interest in the domain of DLs, it is necessary to retrieve both information whose structures need to be extrapolated and the information contained in them. One of the limits of this phase could lie in the creation of the knowledge base (KB) for each object which can have a different structure and present the same information in a different way. Therefore, it will be necessary to pinpoint the present information of interest, defining and outlining it. Finally, we will reconcile the concepts obtained with the two previous steps. For each single metadata found in the first phase, we pinpointed where the information can be found in the metadata representing the knowledge of each object. Starting from this KB, further iterative refining can be made by re-analyzing the information in different phases: with the first phase checking if the information that is not represented by the chosen formalization can be formalized; and a second phase approach analyzing if some information of the Web sites and of the hashtags associated with images on Instagram can be connected. This approach is a combined one, resulting from the combination of the previously mentioned approaches.

The knowledge we want to represent is the one considered of interest by the users in the domain; for this reason, the most important pieces of information and the relevant groupings of hashtags are chosen.

The taxonomy of our project has as its objective to serve as a reference for people and applications with which it integrates: it is necessary that all the involved stakeholders share and recognize those choices and categorizations. Your real goal is to make knowledge manageable, shareable and reusable. Another goal is to implement a Web-based application intended for the optimization of multimedia object metadata classification. Moreover, we want to understand how a resource can be legally used just by looking at the metadata description obtained, by showing the list of the most relevant rights associated with it. Our reference license is called Copy Zero X (www.costozero.org), a customizable license which offers a comprehensive list of intellectual property rights which can be found in the Italian legal system. To classify rights, we created new metadata as DC qualifiers. The information are interesting also as a search criterion, since they would allow for searching only resources with the specific type of rights the user needs.

The basic starting concept is the definition of a KB: in our study, the KB is composed by all kinds of multimedia objects that a DL must manage: ebooks, audiobooks, music, websites, magazines, images and images of Instagram. Through our approach, knowledge is extracted to define a common structure through a taxonomy, in order to classify and make the majority of such knowledge available. We are going to analyze the metadata standards used in multimedia contents management, and define a taxonomy to represent the semantics of these multimedia contents, so that, in turn, the metadata classification can give an unambiguous meaning.

8. Starting from General Concepts

We use standards such as, for example, DC, Exif and the XMP standards, for processing and storing standardized and proprietary information relating to the contents of a file as a starting point of the considered domain. These metadata standards have been described in previous sections. We assume it is possible to use this approach because such standards allow for cataloguing different aspects of multimedia content. With this analysis, we aim at obtaining a complete modeling of the domain of multimedia content properties, together with a uniform representation of the variety of associated metadata.

8.1. Selected Metadata

Metadata belonging to the DC standard are entirely adopted, since they can represent any type of digital resource, due to the generality of the elements semantics. The adoption of the DC standard allows for the system to be OAI compliant, so that the OAI-PMH protocol could be used. The OAI-PMH is a protocol for metadata transmission among systems that requires the use of the 15 core elements of DC in order to be effectively employed. Its added value lies in the higher interoperability the protocol brings into the system, making it a part of a gathering-exchange network of metadata, thus increasing the potential of the final product. On the other hand, since the XMP standard is considerably large, it requires a careful selection not only of its schemas, but also of the metadata included in them. Unlike DC, XMP represents highly specific information, which is not always relevant in the DL context.

The metadata that are considered are thus the ones belonging to the following schemas: XMP basic schema, XMP rights management schema, XMP paged-text schema, XMP Dynamic Media Schema and Exif schema. Among those, only metadata belonging to XMP rights management schema were fully employed, as they represent information on rights associated with the resource. It was also decided to include metadata from MAG 2.0, an application profile specialized in the description of digital resources (derived or born digital). It includes structural and administrative metadata but is lacking in terms of descriptive metadata (it only includes the 15 core elements of DC).

8.2. User-Generated Content with Instagram

Cultural information also exists outside of the institutions that manage the collection of books.

UGC services are designed to provide users with means to support and interpret content. One of our activities involved studying the representation of UGC. Some examples of websites that reached success thanks to this content category are YouTube, Flickr, Twitter, Facebook and Instagram.

The Instagram metadata that we showed in the preceding paragraphs include user data, its navigation and its geographical location. If properly exploited, the metadata, particularly those related to user location, can offer great opportunities and benefits to the various parties involved: both users and DL managers. This large amount is the main cause for the extensive use of a fairly high quantity of metadata. Once all the metadata coming from Instagram had been grouped, the semantics of each and every one of them was evaluated, and, similarly to what was done for DC and XMP, only the most representative and interesting metadata for a DLs were selected.

In addition to the formal content of DLs and metadata, we used hashtags. Tagging is user-generated, user-initiated content, representative of points of engagement between people and catalogs [45]. These points of contact are critical for DLs, for they offer a direct indication of visitor interests, visitor perceptions, and, perhaps, misperceptions. The DLs can learn from watching what and how people use hashtags, perhaps raising points of interest or 'teachable moments' where additional interpretation is necessary. Just as search terms are a direct trace of a trajectory of interest, so too can tags offer an insight into the objects that engage users.

8.3. Mapping

Our next step was the direct mapping between metadata: same meaning, same format, and same data type. We represented their correspondences in a table, so that we could have a clear view of both the metadata we considered in this first phase as a whole, and of how the semantics of the elements overlap. We then chose, where semantics overlap, the ones which were most suited for our purposes. In the resulting table, direct semantic correspondence is represented by placing metadata in the same row, whereas isolated metadata represent a single semantics.

The XMP standard is not in the table because none of its elements has the same semantics as any of the metadata shown above. In Table 1, we showed an example of mapping realized.

Table 1. Mapping.

Dublin Core OAI (Open Archive Initiative) Compliant	Dublin Core Non OAI (Open Archive Initiative) Compliant	MAG (Metadati Amministrativi e Gestionali)	UGC (User-Generated Content) —INSTAGRAM
dc.format	dc.format.extent	mag.metadigit.bib.format	location-id
	dc.format.medium		
dc.identifier	dc.identifier.bibliographicCitation	mag.metadigit.bib.identifier	media-id
		mag.metadigit.stru.element.identifier	
dc.source	-	mag.metadigit.bib.source	-

9. Starting from Specific Concepts

The main part of this phase is based on the interaction between the final user and the DLs: starting from the raw data, we finally define them as subtypes of more general categories up to the root node. This approach is inspired by the atomism concept where the objects we perceive are composed of indivisible units called atoms. The specification of an object in terms of indivisible units and their interactions constitutes the fullest possible description of the object (descriptive aspect), and allows for detection of all the other properties of the object itself (explanatory aspect). Complex objects are instances of the same concept if they are composed by the same kinds of constituents in the same quantity and are connected in the same way.

We have analyzed the specific objects of the domain present in the analyzed portals (images, multimedia objects, documents, ebooks); we then chose tags that we considered as the most suitable for the construction of the taxonomy. We then wondered what information is necessary to retrieve objects in the domain. The tags are then identified as labels that constitute the set of descriptive metadata of a resource.

This analysis is divided into three steps: data collection, grouping, and selection. Data collection has the sole aim to search for multimedia objects that are going to form the reference domain, analyzing and noting the characteristics they possess. Grouping involves dividing the labels collected in the first phase by type. Finally, the selection phase (third phase) consists of isolating those tags that are considered to be the most important to be represented. The amount of occurrences of the characteristics shown in the reference sites, that is, the frequency associated with them, and the possible interest which a DL might have in considering them, are some of the factors taken into account when making the choice.

The websites that were used as reference are: Europeana, Internet Culturale, Cultura Italia, Internet Archive, Open Library, and Project Gutenberg. They use different metadata: Europeana uses DC standard, Internet Culturale applies MAG, and Cultura Italia utilizes the PICO application profile. In the Internet Archive, the metadata collected from the study of the resources are of two types: generic (Title, Author, Publication date) and specific as Last edited By, Last edited dates which identify the author and the date of the last change made. Gutenberg includes ebooks in ePub format and .movi, allowing the download or read online: all books are represented using the same structure, and the same metadata are used for all the resources of the library.

These websites offer a satisfactory overview of the objects that a DL is interested in representing, making it possible to examine and compare the classification of those same objects, as found in portals. After the tags had been collected on six sites, it was necessary to order them. The first step was to list the element types, based on the name assigned to them by the website. Each type of element is associated with one of the following macro-categories: "Image", "Text", "Audio", "Video", "Ebook", "Other". The macro-category "Other" groups together metadata belonging to elements that do not belong to the other labels (such as metadata belonging to the legal documents group from the previous sections). Once the nature of the elements was defined, each group of metadata describing an element was included into the group of metadata identified as strictly related to the nature of that specific element. For example, in a table that has as many columns as there are macro-categories, the tags related to the element "legal documents" (which, as we have already mentioned, belongs to the macro-category "Other) would be all inserted into the same column. This kind of grouping is bound to create semantic duplicates, since the same information could be given different names depending on the site. The strength of this approach that goes from the particular to the general lies in the mechanism to understand how objects are classified and which pieces of information should be selected to represent them. Building a list of tags, divided into macro-categories, is indeed necessary, but, after this step, it is equally relevant to create another list of tags whose semantics serve as a criteria to set them apart, regardless of their names. In order to avoid duplicates, a name that reminds of the tag semantic is assigned; then, the most suitable name is chosen only in a later phase (iteration). For example, the tags "language" and "lingua" (language), found in the elements of Internet Archive and Gutenberg, use different names to refer to the same information; that information will be listed in our table under a single tag.

With a list of metadata by macro-categories, all we have to do is simply choose the tags to keep and those to reject, judging from the frequency of usage on the sites and of the importance of each information for a DL.

10. Comparing the Metadata and Resulting Metadata Taxonomy

We compared metadata from the standards analyzed during the first phase to the data collected during the second phase. The combined approach results from the combination of the previous phases, starting by defining the key concepts and then generalizing and specializing them appropriately. The purpose of the comparison is to verify whether all the features studied during the second phase are represented by the metadata obtained from the first phase. If they are not, new metadata are created, either as an extension of the already chosen metadata (DC allows semantics extensions by adding qualifiers) or as entirely new metadata, creating a new namespace to include them. The process began with a mapping phase, followed by the creation of new metadata and other considerations, such as the representation of ebooks and rights management. Once the iterative phase was completed, and all available metadata were selected, the schema of the taxonomy could be created.

Firstly, it was necessary to perform a comparison between the list of tags and the previously selected metadata, considering their semantics. By doing this, tags whose semantics could not be described by already chosen metadata were identified, so that new metadata could be created for them. To tags whose semantics were similar to a DC element, but more precise, new qualifiers were associated, whereas tags that could not be effectively identified by the DC standard were included in a new namespace called "multimediatype". For instance, the following new qualifiers were created for the DC element "dc.identifier": "isbn", "LoC", "dewey", "iccd". Each one of these qualifiers represent a specific code associated with the digital resource. It is not required to create one metadata for each code type, but it was considered more appropriate to create four qualifiers of "dc.identifier" for the most important codes: International Stardard Book Number (ISBN), Library of Congress Collections (LoC), Dewey, ICCD. For the other codes, the general "dc.identifier" can be used. The code type must also be specified when adding the new metadata. The namespace "multimediatype", instead, includes metadata describing legal documents, publishing information, institutions (for example, museums and libraries), and UGC.

The results of our research showed that there were not any metadata suggesting the optimal software or hardware device for the exploitation of a resource, e.g., an ebook. To overcome this, two new DC qualifiers were created: "dc.format.testedSoftware" and "dc.format.testedDevice". These metadata define the most suitable software and device through which the resource can be exploited. Grey literature can be defined by the level of education of their target user (thus defining the suggested group of users that typically use a specific kind of resource), and the type of document, selected from a list of types that belong to that category (for example, papers, theses and scientific research documents). The metadata are: "dc.audience.instructionLevel" and "multimediatype.documentCategory".

The integration of UGC metadata was performed by focusing on those that had a single semantic during the first phase, and selecting, among them, the most suitable for the DL context. Under the category "multimediatype.ugc", the metadata extracts from Instagram included "user-id" which are associated with "follows" and "followed-by", "media-id" which is associated with "likes", "tag-name", "location-id" in order to represent information about the user who provided the resource by Instagram.

Under the category "multimediatype.ugc", even "mediarestriction", "private", "error" and "statistics" were included, in order to represent information about use restrictions (e.g., some resources can only be used in some countries), the status of the resource (if it is private, only users that have the owner's permission can use the resource), and also information about errors and statistics (such as the average rating or the number of views).

The resulting metadata were used to create the taxonomy structure. The structure has three branches departing from the parent node, related to the main groups of metadata: MAG, DC, and multimediatype. MAG is an application profile with its own structure, so it does not need to be changed, and it could be entirely included in the taxonomy. DC, being composed by simple elements and qualifiers, suggests a further distinction in two levels: one related to simple elements which come directly from the namespace "dc", and the other to the qualifiers of the aforementioned elements, among which the class "Ebook", that comprises metadata "testedSoftware" and "testedDevice" is included. Those metadata, in fact, refer only to that type of resource. Multimediatype metadata can be associated with different kinds of resources with no distinction (those in the "general" category), or to a specific resource. Among those, we include XMP, which consists of the subcategories Audio, Text and Video, and Exif, that includes the subcategory Image. This hierarchy makes it possible to quickly characterize the nature of a resource and the position of the related applicable metadata in the taxonomy during the classification process, and allows for easily selecting the level of detail together with, eventually, the standard to use. The resulting metadata taxonomy is represented in Figure 1.

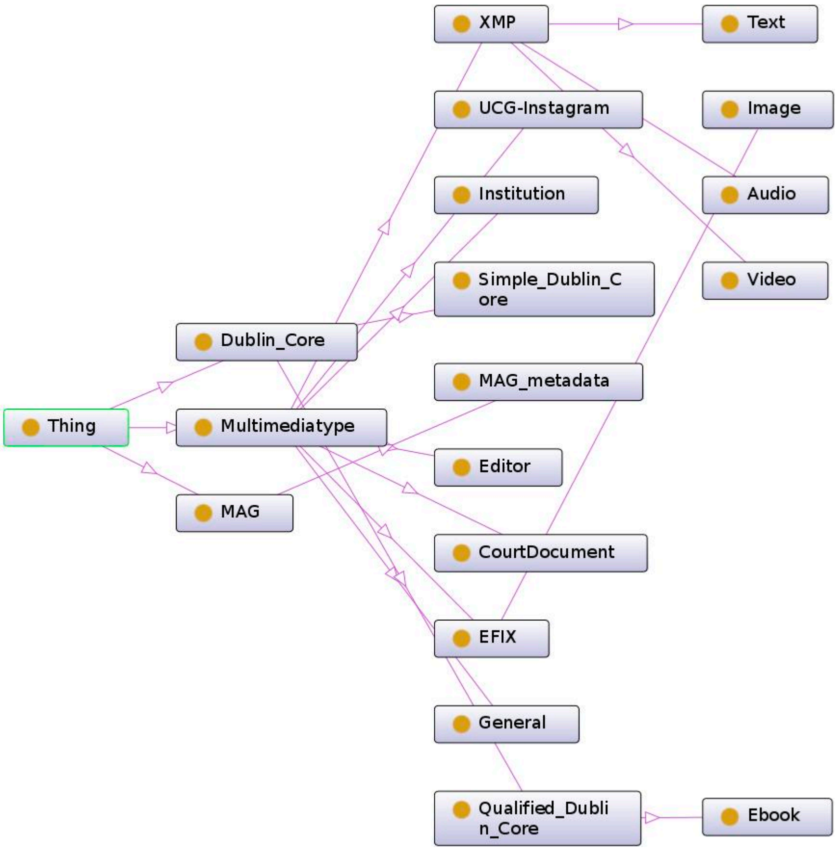

Figure 1. The resulting metadata taxonomy.

11. Discussion

The proposed taxonomy is widely used as it provides a formal representation of the items managed by the library that are explicitly defined and whose content and structure results from an agreement among the stakeholders.

A proper management methodology for the metadata semantic and UGC from Instagram was also proposed, in order to identify entities, priorities, and relationships which are to be found within the domain whereby the system interacts.

The activities planned for the prototype development have been properly scheduled through periodic releases, spaced in time with the aid of "agile" best practices. The final product will be compliant with both national and international standards, such as International Standard Bibliographic Description (ISBD), UNIversal Machine Readable Cataloguing (UNIMARC), ISO 2709, Z39.50 and the ICCU certification, thus proving the capability to exchange information at level 4 of the National Librarian Service (ISBN—Servizio Bibliotecario Nazionale).

Since a validation process for the application is needed, the proposed approach will be extensively tested for effectiveness over the years to come.

The prototype validation process has been planned following the Agile methodology, in order to be flexibly linked to software requirements. To this end, a traceability matrix has been defined, which links each requirement to its corresponding test. The planning process allows for continuous verification of the compliance to requirements during the whole designing phase. Finally, the test phase is to be performed through four steps: unit testing, integration testing, system testing, and acceptance testing of the final prototype.

We answer the previously mentioned research question:

1. It interoperability between metadata and the UGC with Instagram possible?Taking advantage of the metadata provided from Instagram, we selected a set of metadata, with the aim of effectively qualifying UGCs in the context of DLs. The choices were aimed to be as accurate and selective as possible because our objective is to provide a core metadata set for UGCs, so us to ensure that they could easily comply with the specific needs of a specific library.
2. Does using Instagram improves resource management of DLs?UGC of Instagram represents a supporting technology to existing classification systems helping to describe library resources more flexibly and dynamically.

12. Conclusions

Metadata standards are an important mechanism for DLs to manage records and express relationships between them. We focused on interesting information in domain-specific knowledge, thus allowing for the formalization of the metadata that should be currently associated with multimedia objects.

We studied a process to identify existing formalizations and knowledge sources within the domain of DLs, paying attention to multimedia objects and integrating these elements with the content posted by users on the mobile application Instagram.

We found it necessary to introduce two fundamental metadata that can improve the final user experience when he/she wants to quickly and effectively access a DL resource. These two metadata are dc.format.testedSoftware and dc.format.testedDevice: the former suggests a tested application that might be used to easily access the resource, along with some additional information about the operating systems that are compatible with that application; the latter gives some information about the devices which might be used to successfully access the resource (e.g., a specific tablet, smartphone, *etc.*).

The resulting taxonomy, created on the basis of an accurate analysis and the exploitation of widespread standards, provides a descriptive model for the content management in the context of DLs. In particular, resources such as ebooks, which are more and more popular nowadays, need not only to be classified with an exhaustive set of descriptive metadata, but also require specific metadata which makes them easy to use. Moreover, the taxonomy created allows users to select concepts and resources in a simple way. This taxonomy, obtained through the integration of metadata and hashtags associated with the images posted on Instagram with the DL metadata standards, can be used for the interpretation of the most relevant and interesting informative content for a specific user type and in a specific location.

In the future, the taxonomy could be used to provide, in a semi-automated way, search paths associated with different user types or contents. A user that visits a certain DL and publishes a photo on Instagram could automatically receive new suggestions of interesting resources from the application.

Acknowledgments: Simona Ibba gratefully acknowledges the Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Program of the Autonomous Region of Sardinia, European Social Fund).

Author Contributions: Simona Ibba designed the research concept. Both authors jointly conducted analysis, discussed the results and implications, wrote and commented on the manuscript at all stages. Filippo Eros Pani supervised the research activities.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Borgman, C.L. What are digital libraries? Competing visions. *Inf. Process. Manag.* **1999**, *35*, 227–243.
2. Borgman, C.L.; Bates, M.J.; Bates, M.V.; Efthimiadis, E.N.; Gilliland-Swetland, A.J.; Kafai, Y.B.; Leazer, G.H.; Maddox, A.B. Social aspects of digital libraries. In *Background Paper for UCLA—National Science Foundation Workshop*; National Science Foundation Workshop: Arlington, VA, USA, 1995.
3. Hillman, D.I.; Westbrooks, E.L. *Metadata in Practice*; American Library Association: Chicago, IL, USA, 2004.

4. Fox, E.A.; Sornil, O. Digital libraries. In *Encyclopedia of Computer Science*; John Wiley and Sons Ltd.: Chichester, UK, 2003; pp. 576–581.
5. Dalkir, K. *Knowledge Management in Theory and Practice*; Elsevier Butterworth-Heinemann: Oxford, UK, 2005.
6. Houghton-John, S. Library 2.0 discussion: Michael squared. Librarian in Black (blog), 2005. Available online: http://librarianinblack.typepad.com/librarianinblack/2005/12/library_20_disc.html (assessed on 18 November 2015).
7. Peacock, D.; Ellis, D.; Doolan, J. Searching for meaning: Not just records. Available online: http://www.museumsandtheweb.com/mw2004/papers/peacock/peacock.html (assessed on 22 November 2015).
8. Schrier, R.A. Digital librarianship & social media: The digital library as conversation facilitator. *D Lib Mag.* **2011**, *17*, 2.
9. Anfinnsen, S.; Ghinea, G.; De Cesare, S. Web 2.0 and folksonomies in a library context. *Int. J. Inf. Manag.* **2011**, *31*, 63–70. [CrossRef]
10. Ibba, S.; Orrù, M.; Pani, F.E.; Porru, S. Hashtag of Instagram: From Folksonomy to Complex Network. In Proceedings of the 7th International Conference on Knowledge Engineering and Ontology Development (KEOD), Lisbon, Portugal, 12–14 November 2015. [CrossRef]
11. Hu, Y.; Manikonda, L.; Kambhampati, S. What We Instagram: A First Analysis of Instagram Photo Content and User Types. In Proceedings of the 8th International AAAI Conference on Web and Social Media (ICWSM), Ann Arbor, MI, USA, 1–4 June 2014.
12. Furini, M.; Tamanini, V. Location privacy and public metadata in social media platforms: Attitudes, behaviors and opinions. *Multimedia Tools Appl.* **2015**, *74*, 9795–9825. [CrossRef]
13. Highfield, T.; Leaver, T. A methodology for mapping Instagram hashtags. *First Monday* **2014**, *20*. [CrossRef]
14. Schwartz, D.G. The Emerging Discipline of Knowledge Management. *Int. J. Knowl. Manag. IGI Glob.* **2005**, *1*, 1–11. [CrossRef]
15. Lagoze, C.; Van de Sompel, H. The Making of the Open Archives Initiative Protocol for Metadata Harvesting. *Libr. Hi Tech* **2003**, *21*, 118–128. [CrossRef]
16. Adobe Systems Incorporated, Adobe XMP Specifications, Additional Properties. 2010. Available online: http://www.adobe.com/content/dam/Adobe/en/devnet/XMP/pdfs/XMPSpecificationPart2.pdf (assessed on 18 September 2015).
17. Technical Standardization Committee on AV IT Storage Systems and Equipment. *Exchangeable Image File Format for Digital Still Cameras: Exif Version 2.2*; Standard of Japan Electronics and Information Technology Industries Association: Tokyo, Japan, 2002. Available online: http://www.exif.org/Exif2--2.pdf (assessed on 21 October 2015).
18. Stamatios, G.; Tsapatsoulis, N. Instagram Hashtags as Image Annotation Metadata. In *Artificial Intelligence Applications and Innovations*; Springer International Publishing: Berlin, Germany, 2015; pp. 206–220.
19. Danowski, P. Library 2.0 and user-generated content: What can the users do for us. In Proceedings of the World Library and Information Congress: 73rd IFLA General Conference and Council, Durban, South Africa, 19–23 August 2007.
20. Ulli, W.; Mehler, A.; Heyer, G. Towards Automatic Content Tagging-Enhanced Web Services in Digital Libraries using Lexical Chaining. In Proceedings of the Fourth International Conference on Web Information Systems and Technologies, Funchal, Madeira, Portugal, 4–7 May 2008; Volume 2.
21. McNulty, N. The development of a user-generated digital library: The case of the Ulwazi programme. In Proceedings of the IST-Africa Conference and Exhibition (IST-Africa), Nairobi, Kenya, 29–31 May 2013; pp. 1–11.
22. Chih-Ming, C.; Yong-Ting, C.; Chin-Ming, H.; Chin-Wen, L.; Chia-Meng, H. Developing a Taiwan library history digital library with reader knowledge archiving and sharing mechanisms based on the DSpace platform. In *The Electronic Library*; Emerald Group Publishing Limited: Bingley, UK, 2012; Volume 30, pp. 426–442.
23. Jin-Cheon, N.; Tun Thura, T.; Arie Hans, N.; Fauzi Munif, H. A Sentiment-Based Digital Library of Movie Review Documents Using Fedora/Une bibliothèque numérique de documents critiques de films basée sur les sentiments en utilisant Fedora. *Can. J. Inf. Libr. Sci.* **2011**, *35*, 307–337.
24. Witten, I.H.; Bainbridge, D.; Paynter, G.; Boddie, S. *Importing Documents and Metadata into Digital Libraries: Requirements Analysis and an Extensible Architecture*; Springer: Berlin, Germany, 2002; pp. 390–405.

25. Haslhofer, B.; Klas, W. A Survey of Techniques for Achieving Metadata Interoperability. *J. ACM Comput. Surv. CSUR* **2010**, *42*. [CrossRef]
26. Chan, L.M.; Zeng, M.L. Metadata interoperability and standardization—A study of methodology, Part I: Achieving interoperability at the schema level. *D Lib Mag.* **2006**, *12*. [CrossRef]
27. Stadlhofer, B.; Salhofer, P.; Durlacher, A. An Overview of Ontology Engineering Methodologies in the Context of Public Administration. In Proceedings of the 7th International Conference on Advances in Semantic Processing, IARIA, Porto, Portugal, 29 September–3 October 2013; pp. 36–42.
28. Salembier, P.; Sikora, T.; Manjunath, B.S. *Introduction to MPEG-7: Multimedia Content Description Interface*; John Wiley & Sons, Inc.: New York, NY, USA, 2002.
29. Martinez, J.M.; Koenen, R.; Pereira, F. MPEG-7: The Generic Multimedia Content Description Standard, part 1. *IEEE Multimedia* **2002**, *9*, 78–87. [CrossRef]
30. Suárez-Figueroa, M.C.; Carmen, M.; Atemezing, G.A.; Corcho, O. The landscape of multimedia ontologies in the last decade. *Multimedia Tools Appl.* **2013**, *62*, 377–399. [CrossRef]
31. García, R.; Celma, Ò. Semantic Integration and Retrieval of Multimedia Metadata. In Proceedings of the 5th International Workshop on Knowledge Markup and Semantic Annotation, Galway, Ireland, 7 November 2005; pp. 69–80.
32. Hunter, J. Adding Multimedia to the Semantic Web—Building an MPEG-7 Ontology. In *Multimedia Content and the Semantic Web: Standards, Methods and Tools*; Stamou, G., Kollias, S., Eds.; John Wiley & Sons Ltd.: New York, NY, USA, 2001; pp. 75–100.
33. Scherp, A.; Eißing, D.; Saathoff, C. A Method for Integrating Multimedia Metadata Standards and Metadata Formats with the Multimedia Metadata Ontology. *Int. J. Semant. Comput.* **2012**, *6*, 25–49. [CrossRef]
34. Gartner, R. *METS: Metadata Encoding and Transmission Standard*; JISC Techwatch report TSW; Oxford University Library Services: Oxford, UK, 2002.
35. Denise, D.M. NISO Standard Z39.7. The Evolution to a Data Dictionary for Library Metrics and Assessment Methods. *Ser. Rev.* **2004**, *30*, 15–24.
36. MAG, Comitato. *Metadati Amministrativi e Gestionali: Manuale Utente*; Pierazzo, E., Ed.; ICCU: Roma, Italy, 2006.
37. Kogalovskii, M.R.; Parinov, S.I. The taxonomy of semantic linkages of information objects in research digital library content. *Autom. Doc. Math. Linguist.* **2015**, *49*, 163–171. [CrossRef]
38. Gonçalves, M.A.; Fox, E.A.; Watson, L.T. Towards a digital library theory: A formal digital library ontology. *Int. J. Digit. Libr.* **2008**, *8*, 91–114. [CrossRef]
39. Spiteri, L.F. The structure and form of folksonomy tags: The road to the public library catalog. *Inf. Technol. Libr.* **2007**, *26*. [CrossRef]
40. Peterson, E. Beneath the metadata: Some philosophical problems with folksonomy. *D-Lib Mag.* **2006**, *12*. [CrossRef]
41. Lee, H.-J.; Neal, D. A new model for semantic photograph description combining basic levels and user-assigned descriptors. *J. Inf. Sci.* **2010**. [CrossRef]
42. Avery, J.M. The democratization of metadata: Collective tagging, folksonomies and web 2.0. *Libr. Stud. J.* **2010**, *5*, 8.
43. Ding, Y.; Jacob, E.K.; Fried, M.; Toma, I.; Yan, E.; Foo, S.; Milojević, S. Upper tag ontology for integrating social tagging data. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 505–521. [CrossRef]
44. Pani, F.E.; Lunesu, M.I.; Concas, G.; Baralla, G. The Web Knowledge Management: A Taxonomy-Based Approach. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management*; Fred, A., Dietz, J.L.G., Liu, K., Filipe, J., Eds.; Springer Verlag: Berlin, Germany, 2015; pp. 230–244. [CrossRef]
45. Trant, J. Tagging, Folksonomy and Art Museums: Early Experiments and Ongoing Research. Available online: https://journals.tdl.org/jodi/index.php/jodi/article/view/270/277 (accessed on 10 September 2015).