This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

Foundations of Digital Universities

Abstract. Universities need to maintain data about various resources. They include intellectual creations that are the focus of (Digital) Libraries, but also courses, research projects and people. Data about the first are maintained in institutional repositories, while data about the others in various databases designed to support specific vertical applications. In this paper, we propose a uniform treatment of such resources leading to Digital Universities, i.e. a set of key resources, methodologies and tools appropriately organized to effectively support universities' users. This requires new methodologies, data models, authority control mechanisms and system infrastructures able to support a broader range of services.

Keywords. Digital libraries and archives, decision support systems, digital universities, metadata, authority control

1. Introduction

Universities are institutions of higher education and research which grant academic degrees in various subjects. The services they offer are pivoted on the production, custodianship, fruition and dissemination of knowledge. Knowledge is embedded in a few types of fundamental key assets which include intellectual creations - such as papers, books, thesis and patents - as well as other entities such as courses and research projects. Similarly, people can be categorized in knowledge producers (e.g. researchers), knowledge administrators and holders (e.g. librarians and professors) and knowledge consumers (e.g. students), each of them being individually characterized by different expertise, skills and duties.

The approach followed so far by universities to offer knowledge-centric services to their users is to put in place a complex ecosystem of libraries, digital libraries and IT systems, each of them engineered to serve a specific vertical application. A vertical application supports a specific business process and targets a smaller number of users with specific skill sets and job responsibilities within an organization[1]. For instance, library management systems are used by the university libraries to track items owned, items borrowed, orders made and bills paid; OPAC systems provide online access to catalogues of library material such as books and thesis; institutional repositories are used to archive digital material such as conference papers and journal publications. Moreover, applications also include institutional websites, human resources (HR) management, teaching support, project management, and knowledge transfer IT systems. The various needs are met by applying standard software engineering approaches that lead to the development of systems tailored for the specific application they need to serve, or to the adoption of the most appropriate off-the-shelf system addressing the identified business need.

This approach has been quite successful so far, in that each system is engineered and maintained separately. Each system focuses on a small set of functionalities and knowledge assets for a homogeneous group of users. The responsibility is confined to people with specific skills and duties. This separation of data and duties makes the engineering process easier and the management of resources effective.

On the other hand, the main drawback of this approach is that data about the key assets is scattered across multiple separate information silos, data is often duplicated and difficult to correlate due to the diversity in format, metadata, conventions and terminology used. This situation reaches extreme consequences when different departments within the university decide to use different IT systems to manage their own assets. As a result, such *data fragmentation* and *data diversity* makes *entity search* [8] and *data analytics* [3][9] very challenging to be addressed globally [19]. The capacity to correlate and search data about the key assets at the level of the whole university is vital to support effective discovery, visualization and navigation services to universities' users as well as to support the university governance.

Consider for instance the institutional website of the university. It presents university facilities, as well as faculty members together with their publications, projects and courses. Examples of universities that provide such kind of integrated services are the University of Toronto[2] and the University of Hong Kong[3]. Offering these services requires the capacity of collecting all the necessary data about the key

[1] http://www.webopedia.com/TERM/V/vertical_application.html
[2] https://focus.library.utoronto.ca/
[3] http://hub.hku.hk/

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

entities scattered across the various information silos, and to make their representation homogeneous in terms of schema and terminology used to describe them. In case the institutional website needs to be offered in multiple languages, it also requires the disambiguation of the meaning of the terms used to describe the various entities and their translation in all the necessary languages.

Other services are centered on the capacity of universities to periodically census and evaluate the quality of the work done. In Italy for instance, universities and research institutes must provide every year to the National Agency for the Evaluation of Universities and Research Institutes (ANVUR) detailed information about the outcomes of the research activities they conducted. It includes (to mention a few) information about publications (their number, type and full citation), projects acquired (their name, local coordinator, budget, and sponsors), patents produced, awards won, and public engagement activities (complete list of public events organized with attendance information). This is requested in order to assess the quality of research conducted by each Italian institution. Within each institution, this is typically done by hand by administrative and technical staff of several offices spanning all academic departments. In fact, the necessary data is stored in different repositories, if not in the PC of individual people. This data collection process takes months with a huge cost in terms of human resources.

A few initiatives recently provided solutions to universities having such needs. They include the *Linked Universities* [4] and the *VIVO* [5] [26] initiatives. They rely on standard Semantic Web technologies and tools to extract data from the available information silos, to convert it in RDF and store it in a triple-store. Data is then retrieved by using the SPARQL query language. Their implementations vary according to local information provider's policy and practice. These solutions are technologically advanced but, as we extensively describe in the paper, methodologically weak. In fact, very little methodological and technological support is given by these initiatives for data curation.

Conversely, (digital) libraries have a strong tradition in data and metadata curation and are indubitably methodologically very strong, but are limited in the scope. In fact, data fragmentation and data diversity are typically addressed in libraries by ensuring the adoption of standard data models and authority control practices that ensure effective resource discovery and interoperability between libraries [11]. Repositories are actually designed to adhere to these standards and practices. However, current data models and discovery services [7], being them centered on intellectual creations, do not account for all the key relevant entities and therefore cannot accommodate for all the services required by universities' users. Moreover, they do not offer a solution to integrate data coming from such heterogeneous systems.

In order to appropriately tackle these difficulties, we envision that the universities of the future should leverage on Digital Universities, i.e. a set of key resources, methodologies and tools appropriately organized to effectively support universities' users. In particular, the methodologies and tools need to guarantee for a high level of data quality. In essence, Digital Universities are for a university what Digital Libraries are for a library or museum. This requires new methodologies, data models, authority control mechanisms and platforms able to support a broader range of entity types and services. By authority mechanisms we mean methodologies, rules and supporting tools to control the form of entity names and the terminology used to describe such entities.

In infrastructural terms, the solution we propose is centered on the *Knowledge HUB*, a new platform that goes far beyond the capabilities of a standard institutional repository. The Knowledge HUB supports appropriately skilled university staff in the definition of a centralized data model and controlled vocabularies that - through the setup of appropriate authority control mechanisms - allow the cataloguing and the classification of all the key knowledge assets of the university. The fragmented and heterogeneous data about the key assets that are maintained separately by the various legacy systems of the university is appropriately extracted and processed to comply with the centralized data model and controlled vocabulary of the HUB. Thus, the Knowledge HUB acts as a trusted proxy to the legacy systems in that it offers the basic facilities to (a) homogenize and integrate data coming from the data sources, (b) keep the content of the central catalogue aligned with the data sources, and (c) support centralized discovery and interoperability services. The data model and the controlled vocabularies of the HUB will have to be designed to accommodate for all the entities and the terminology necessary to support the centralized services. Given that these services can vary in time, the HUB will have to support the dynamic extension of the data model and the controlled vocabularies to accommodate for new emerging needs.

This should be also reflected in the universities' organization in that the Knowledge HUB should be under the responsibility of a single division that is in control of the data management and governance functions, as well as the technologies employed to support them. We believe that the competences required to manage the Knowledge HUB include and extend those of library and information scientists and should converge towards a new professional that somebody already calls the *data scientist* [15]. We are at

[4] http://linkeduniversities.org/

[5] https://wiki.duraspace.org/display/VIVO

the beginning of a new era that will be based on the pervasive usage of data. This will deeply change modern society towards a smarter society [14], as it will encompass not only a technological but also a social change. Universities are already part of this move as they will progressively provide their services in a blended fashion in a virtuous integration between the real and the virtual world.

The rest of the paper is organized as follows. Section 2 "State of the Art" describes the state of the art in Digital Libraries and Semantic Web communities. Section 3 "Digital Universities" introduces the concept of Digital Universities as natural extensions of Digital Libraries, it presents key knowledge assets and services, and the system infrastructure required. Section 4 "The methodology" describes the sequential steps that need to be followed each time a new centralized service is required by university users. These steps are described in more detail in Sections 5-9. Section 10 "Trento as Digital University" explains the first steps that we are moving at the University of Trento in Italy towards implementing this vision via the development of a system infrastructure and the first centralized services. Section 11 "Conclusions" concludes the paper by summarizing the work done and the next steps.

2. State of the Art

Libraries and Digital Libraries

In libraries, in museums, or in archives, a catalogue is a collection of organized data describing the information content managed by an institution [5]. Cataloging is the process that, guided by rigorous rules, information scientists follow to create and maintain metadata in order to effectively represent and exploit the information content.

The entities traditionally at the center of library cataloging are the intellectual and artistic creations [7]. They play a privileged role in that they are those returned by the search facilities. Several data models have been proposed and adopted in time. The strengths and weaknesses of each model are continuously under scrutiny by the community [2]. The evolution of standard models, and the cataloging rules governing them, reflects the availability of new technologies and the emergence of new needs, such as the necessity to account for digital objects, or for new entity types and properties, or new modelling paradigms [1].

Authority control makes sure that each entity is assigned a unique header such that it can be uniquely identified and referred to [7]. This includes the curation of identifiers and names. Entities of different types are maintained in authority files, i.e. repositories of homogeneous entities such as authors, locations and organizations. Vocabulary control enforces unique headers for subjects [6], thus making sure that there is an unambiguous way to refer to each subject. Current methodologies employed in library science rely on thesauri providing standard terms for subjects arranged hierarchically from broader to narrower terms [11].

Altogether, the adoption of standard data models, authority control rules, and vocabulary control ensure high levels of data quality. In addition, knowledge organization systems, such as subject classifications, can be employed to effectively organize, browse and search intellectual and artistic creations.

Though methodologically very strong, libraries' approaches cannot be directly applied to universities given their narrow scope. In fact, they only focus on intellectual and artistic creations and this is reflected in the services they support. As observed, universities needs to maintain and offer services centered on a broader range of entities. Also, they need to deal with a high fragmentation and diversity of data that is avoided in libraries though the adoption of standard data models and data exchange protocols.

Semantic Web solutions to university scenarios

Semantic Web technologies describe and store data about the relevant entities using formal languages, such as RDF and OWL. In such languages, the entities are described in terms of subject-property-object triples linked with each other via URIs, i.e. persistent identifiers that allow to uniquely identifying a resource in the Web. In this way, triples can be stored in multiple physically distributed repositories and still refer to each other. Altogether, the defined triples form a (distributed) *knowledge graph* where the nodes represent entities and the arcs between them represent their properties. Such models, by representing data in a uniform machine-readable format *with explicit meaning*, support the development of intelligent interconnected services able to search and navigate beyond the limits of a physical repository and identify information about the same entity spread all over the Web. Moreover, these models are "open" in nature in that they support the definition of new properties *at operating time* without affecting the already defined data and services exploiting them.

At the best of our knowledge, there are a few Semantic Web initiatives aiming at providing support to universities interested at developing and maintaining their own institutional knowledge graph.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

The *Linked Universities* initiative is an alliance of European universities engaged into exposing their public data as linked data. In the spirit of Linked Open Data, its aim is to encourage universities in delivering their data in RDF such that they can be linked with each other, thus promoting interoperability between them. Institutions participating to the initiative include the Open University [24] and the Ege University [25]. They rely on standard Semantic Web technologies and tools to extract, convert and store data in RDF, as well as to query it using the SPARQL query language. For far, there is no evidence of coordination among the universities participating to the initiative to converge to a common data model and supporting platform.

The VIVO [26] initiative provides facilities to universities to publish their data in RDF. The VIVO Harvester Java library can be used by programmers to extract data from original sources and transform it in RDF. RDF data is stored in a triple-store. VIVO provides a default OWL ontology to describe entity classes and properties which are of interest to a university. The OWL ontology can be extended by means of a graphical editor. The VIVO knowledge graph can be accessed via VIVO APIs that allow issuing SPARQL queries to the triple-store. VIVO implementations vary according to local information provider policy and practice.

We understand these two initiatives as pure Semantic Web and Linked Data applications. In fact, their main purpose is the publication of data in RDF to support interoperability between universities. Actually, VIVO also offers basic templates that can be used to develop institutional websites as main service to university users.

Very little support is given by these initiatives for data curation in the way in which this is understood in libraries. In terms of authority control, URIs play the role of unique headers, though nothing prevents an entity to have multiple URIs across datasets. Duplicates are handled at importing time by discovering and linking them via the *owl:sameAs* property. For instance, both VIVO and the Ege University solutions rely on String similarity, especially applied to person names. Programmers will have to implement their own solutions to discover them. This approach is not only weak - as it does not rely on any identifier and name authority rules - but it also means that duplicates are not merged. As a result, multiple equivalent entities can be returned by the queries and applications will have to appropriately reconcile them before exploiting and visualizing the results. No support is directly provided in terms of vocabulary control, in that vocabularies need to be defined elsewhere. In fact, they only rely on standard linking mechanisms based on URIs to link terms to external vocabularies. They do not seem to provide any facility or suggest any methodology to control and enforce terminology. VIVO provides a simple mechanism to handle data provenance which is attached to entire sub-graphs. We believe this level of granularity cannot be sufficient in that properties of the same entity may come from different sources. Understanding the provenance of each piece of information is fundamental to assess the level of authority of data.

3. Digital Universities

We define a digital university as *a set of key resources, methodologies and tools appropriately organized to effectively support universities' users*. Digital Universities are natural extensions of Digital Libraries. We leverage on the strengths of the methodologies employed in libraries and extend their scope to all those entities and services required by universities. The broader scope requires new (or extended) methodologies, data models, cataloging, authority control mechanisms and system infrastructures. In particular, they require the capacity to cope with the *data fragmentation* and *data diversity* which is intrinsic in university settings, due to the size and the variety of systems that are employed across the various departments.

Knowledge assets and services

The key knowledge assets of a university include intellectual creations, courses, research projects, people and facilities. The university services which are centered on these entities include:

- ***Discovery services:*** these services leverage on basic search facilities to give the opportunity to issue expressive queries asking for any kind of entity on the basis of any of their properties [8]. Table 1 provides examples of queries and corresponding system capabilities required. Dedicated knowledge browsers should be developed to support users in searching and navigating data in tabular, hierarchical (e.g. subject classifications) and other graphical visualization modes.
- ***Big Data analytics services:*** they support institutions in the decision-making processes. Table 2 provides examples of analytics and corresponding system capabilities required.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

- ***Institutional services:*** they exploit knowledge content to offer innovative ways to present institutional information to different actors. They play a crucial role in communication. For instance, these services support the capacity of the university to present uniformly and consistently information across different institutional websites. For instance, the same information can be published on the main institutional website of the university, the website of a specific department or of a specific professor. Transformation procedures may take care of adapting data in terms of schema, language, terminology and granularity of information according to the purpose and audience.
- ***Interoperability services:*** these services support the mapping and import/export of data from/to existing standards. An example of service of this kind is the automatic centralized publication of the institutional Linked Open Data that can be exchanged with other universities or research institutes. In compliance with the institutional and national regulations governing privacy and Intellectual Property Rights, they should support the conversion of the data in an appropriate standard model (e.g. the BIBFRAME [13]) and syntax (e.g. SKOS, RDF or JSON) linked with standard vocabularies. These services support the capacity to answer queries across multiple universities. For instance, in the future this should support students in the search for educational opportunities across all universities in the world. Via these services, data can be also transferred to the national government, in case it asks for periodic evaluations of the quality of research, or to companies for knowledge transfer purposes.

Table 1. Examples of queries

Query	Capability required
Give me the list of publications written by John Doe with CC0 access rights	This is a classical query in libraries, being it centered of intellectual creations and their direct properties.
Give me all the courses taught by John Doe	It requires the capability to query for entities different from intellectual creations.
Give me the names of the people who both teach at least a course and are coordinators of a research project	It requires the capability to nest or unify queries.
Give me the list of the top 10 most productive persons in terms of projects	It requires the capability to perform complex operations such as the sum of certain properties, i.e. the budget, and return results in descending order.

Table 2. Examples of analytics

Analytics	Capability required
Give me the trend of publications about the subject *finance* over the past 5 years	This is a classical need in libraries, being it centered of intellectual creations and their subject. The service is needed to discover if the university has enough experts in the subject; if not, the governing body may decide to open new calls to hire experts.
Give me the percentage of publications in open access with respect to the total	This is a classical need in libraries, being it centered of intellectual creations and their subject. The service is needed to understand if the university is sensitive enough to open access; if not, appropriate campaigns can be launched to promote a change in the publishing culture.
Give me the trend of funded projects over the past 5 years	This requires the capacity to deal with entities different from intellectual creations. The service is needed to discover if the university has enough capacity to attract funds; if not, the governing body may decide to hire experts that can assist researchers in the preparation of the project proposals.

The system infrastructure

In order to effectively support the university services, despite the initial data fragmentation and diversity, Digital Universities require a new system infrastructure, exemplified in Figure 1. It supports appropri-

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

ately skilled university staff in the data management. The core of this infrastructure is what we call the *Knowledge HUB* in that it acts as a trusted proxy in between the services and the legacy systems. A fundamental advantage of relying on a central Knowledge HUB is the possibility to reuse the same data - represented in a uniform data model and terminology - for multiple services, with obvious advantages in terms of maintenance and costs of such services.

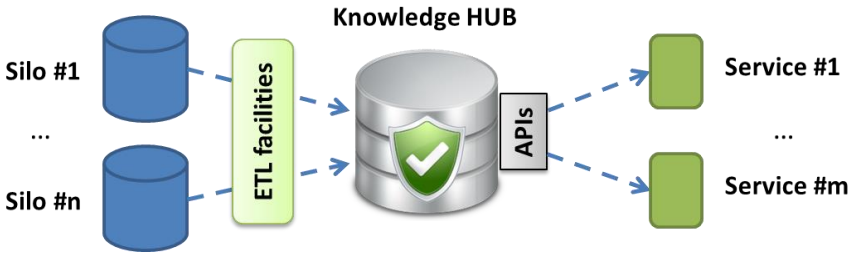

Figure 1 - The Digital Universities system infrastructure

The system infrastructure will have to fulfill the following requirements:

- ***Provide centralized access to information.*** The system infrastructure, through the Knowledge HUB, should offer centralized access to information that is originally stored in different information silos, maintained by legacy systems, and codified following different data models and formats. This is necessary to make sure that legacy systems can continue to function as usual, thus benefitting from all the advantages that come from their vertical applications. Advantages include contained cost, dedicated business processes, focuses data, dedicated users and confined responsibilities. At the same time, data about all the key entities that is necessary to support centralized services should be identified and replicated (or better, indexed) in the Knowledge HUB. Addressing this challenge requires the availability of data extraction, transformation and load (ETL) facilities, as well as the capacity of the HUB to correlate and merge data about the same entity. In particular, merge facilities are essential to avoid the presence of duplicates.
- ***Support the definition of the data model.*** A new standard conceptual data model is required for universities. The main purpose of the standard is to provide a common core of entity types and properties to favor interoperability between universities. It should cover a broader range of entity types w.r.t. those traditionally maintained in Digital Library catalogues. Given the good practices of libraries to stick to well-established data models [3][4], the model should be designed as extension of, or at least as aligned with, existing standards. Nevertheless, the different institutional needs demand for the capacity of the system to support the customization and extension of the standard data model as required by the centralized services that in time will be required by a certain university. The model should also accommodate for provenance, reputation, versioning, intellectual property rights (IPR), licensing, privacy and access control.
- ***Support the setup of the authority mechanisms.*** Authority control is fundamental to guarantee that the catalogued data is of adequate quality [5]. The system should support the setup of the name authority (to regulate the form of names), the identity management (to guarantee for the uniqueness of the entities), and the vocabulary control (to standardize the terminology used) mechanisms. All together these mechanisms allow data to be appropriately governed.
- ***Support the development of the services.*** The Knowledge HUB should provide Application Programming Interfaces (APIs) to support the development of university services such that they can all query the HUB and exploit the same content.

4. The methodology

Governing a Digital University will require the application of an adequate working methodology. We preliminary propose the following approach constituted by 7 sequential steps to be followed each time a new centralized service is required by university users. The methodology ensures that the system infra-

structure is incrementally adapted to support the new service. The adaptation must ensure that services already supported continue to function as expected.

1. ***Collecting service requirements:*** the first step consists in collecting the requirements of the new service, in terms of functionalities, target users, and data required.
2. ***Setting up the data model:*** we assume a core standard data model exists in the Knowledge HUB, even when no service has been developed yet, constituted by entity types and properties necessary to describe typical key university knowledge assets. The data model is eventually extended incrementally with entity types and properties which are necessary to support the new service.
3. ***Setting up the authority control mechanisms:*** identifiers and name authority rules, and the tools necessary to enforce them, are setup for all the entity types necessary to support the new service.
4. ***Setting up the controlled vocabularies:*** controlled vocabularies are incrementally extended to accommodate for the new terms required to describe the new entity types and properties.
5. ***Data hunting:*** existing university systems are assessed in order to identify possible sources for the data necessary to sustain the service. The following cases can arise: (a) there is only one system that can provide the necessary data, (b) multiple systems, possibly maintained by different departments, can provide part of the necessary data, that can eventually partially overlap or even be in conflict, or (c) existing systems cannot provide all the necessary data. In the latter case, it is necessary to develop new systems able to complement the missing data.
6. ***Populating the knowledge HUB:*** by means of extract, translate and load (ETL) tools data is extracted from the available systems, translated according to the defined data model and the controlled vocabularies, and loaded in the HUB. In particular, the role of the data model, authority control and vocabularies is to homogenize data by means of a uniform schema and terminology, thus resolving the initial fragmentation and data diversity. This process requires an adequate infrastructure able to semi-automate the process and to keep the HUB aligned with the sources (e.g. once a day). An example of case in which human intervention is required is to fix mistakes in the data or to accommodate for missing terms or synonyms in the controlled vocabularies.
7. ***Implementing the service:*** the service is implemented. It uses the Knowledge HUB as data source by means of its APIs.

In the following we describe in more detail the steps from 2 to 6 of the proposed methodology. We assume that the services to be developed (step 1) are those presented in Table 1 and Table 2. We skip the step 7 as it depends of the specific technology employed. To make easily digestible the various concepts presented, we use simple data structures for the data representation.

5. The Data Model

The core data model needs to accommodate for all the key assets of universities. The entity types required include intellectual creations such as papers, books, thesis and patents, but also other entities such as courses, research projects, and university departments.

Figure 2 provides an example of object-oriented data model. Rounded boxes represent entity types; for each entity type, the attributes are listed in the boxes with an associated data type; squared brackets indicate that the attribute allows for multiple values; arrows between boxes represent relations. The arrows marked with *extends* codify inheritance of properties from a parent entity type. For instance, the entity type Project inherits start and end dates from Event.

The following attributes are associated to all the entities (see Entity in Figure 2):

- ***Class.*** The class attribute is used to specify the kind of entity as more specific than the entity type, thus avoiding the specification of additional properties. Each entity is associated exactly one class. For example, Trento can be defined as Location, by specifying the values of the properties foreseen for locations, and by specifying that it is of class City. This attribute is important to limit the amount of entity types to be defined.
- ***Name.*** Each entity is associated at least one name. FRAD [5] underlines the necessity to accommodate for name variants (e.g., Fausto Giunchiglia and F. Giunchiglia) and variations (e.g., Fausto Pietro Giunchiglia) in multiple languages (e.g. Faust Giunchiglia). Therefore, names require appropriate data structures. In the model in Figure 2, this important requirement is addressed by introducing the data type Name (see Section 7 “Authority Control”).

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

- ***Identifier.*** Each entity is associated one or more identifiers. Multiple identifiers can be defined for an entity type. They are represented with underlined font in Figure 2. There cannot be two entities which are instances of the same entity type with the same value for an identifier (see Section 7 "Authority Control").

- ***Subject***. While in libraries the only entities that have a subject are the intellectual and artistic creations [6], in university settings all the entities can be potentially associated with one or more subjects. For instance, courses are units of teaching led by one or more instructors (teachers or professors) centered on one or more subjects; research projects investigate specific subjects; people are experts of specific disciplines and subjects. Subjects allow framing the university in sets of knowledge assets.

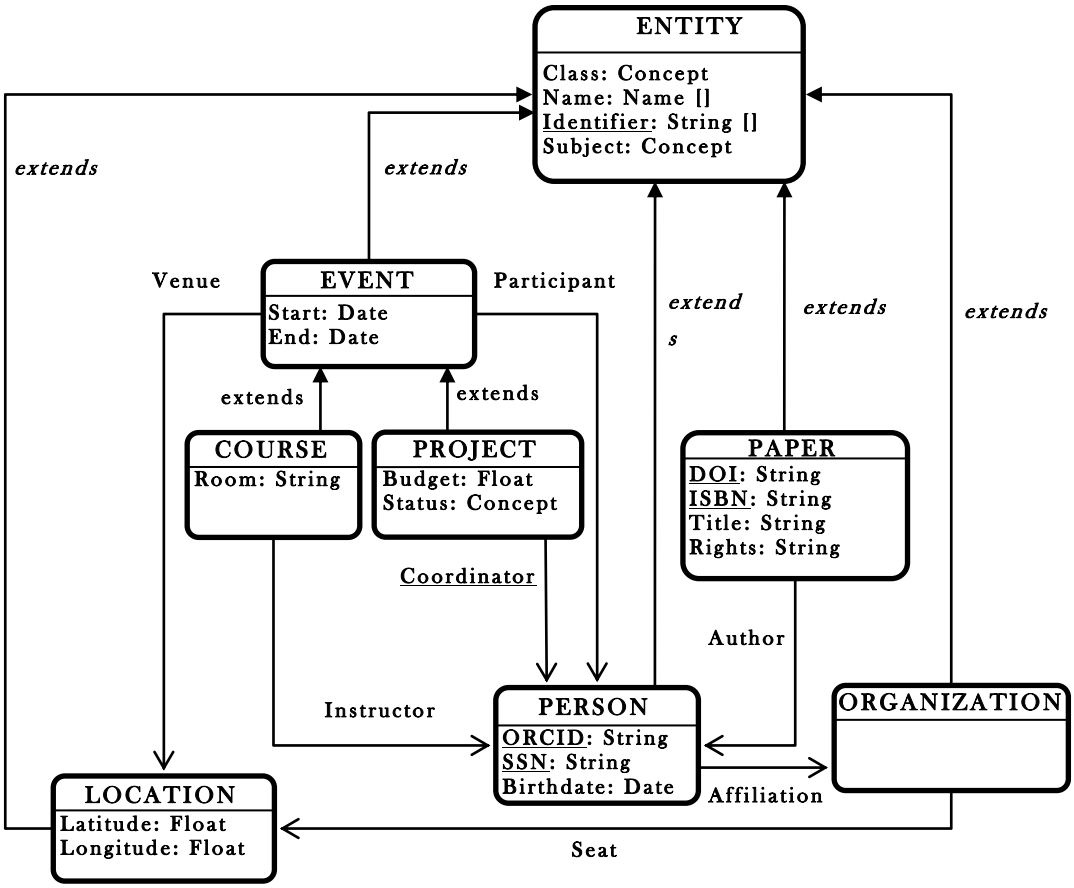

Figure 2 – An example of data model

Two important aspects about the data model are underlined here:

- ***Some of the attribute values are controlled.*** The data type Concept is used to denote those attributes whose values are controlled (see Section 8 "Vocabulary Control). In particular, both the class and subject attributes are controlled in their possible values.

- ***None of the entity types plays a privileged role*** in the data model given that any of them can be returned by queries or be exploited by a service.

6. Authority Control

Libraries *co-locate* information about the same entity [5], i.e. they put in the same place (logically or physically) information about the same individual object described. This is accomplished by establishing unique headers or controlled access points that, following precise rules and formats, support their *identification* and *reference*. As the terms suggest, *identification* consists in providing mechanisms that allow the identification of a specific entity by distinguishing information about it from information about the other

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

entities in the catalogue; *reference* consists in providing mechanisms that allow information about a certain entity to be referred within the description of another entity, for instance to describe a relation between two entities (e.g. person A is the author of paper B). Controlled access points can rely on names, identifiers or combinations of those.

The capacity to appropriately govern controlled access points for identification and reference is what we call *identity management*. For instance, in database technologies identification is supported via the definition of primary keys, while reference is supported via foreign keys. Both primary keys and foreign keys can be defined as combinations of multiple entity properties. In Semantic Web technologies, URIs provision both identification and reference. This means that while databases technologies can support controlled access points of any kind, Semantic Web technologies typically employ URIs. As a matter of fact, in triple-stores duplicates are tolerated and typically linked with each other via the *owl:sameAs* property. It is typically the user who is in charge to discover duplicates and explicitly define such property. This is a rather important limitation that can be partially overcome by simulating primary keys and unique constraints via *owl* constructs [27][28], via combinations of inverse property function and minimum cardinality constructs. Violations of these constraints are detected by reasoners that, however, do not guarantee consistency of transactions.

In order to overcome the current limitations of Semantic Web technologies, we suggest that *identification* should be supported by the Knowledge HUB via the definition in the data model of multiple identifiers that correspond to different controlled access points. An identifier may correspond to one property (e.g. the DOI for publications and ORCID for people) or a set of properties (e.g. the combination of name and coordinator for a project). Multiple identifiers can be defined for the same entity type. For instance, in Figure 2, ORCID and Social Security Number (SSN) are defined as identifiers for people; DOI and ISBN are defined as identifiers for publications. Identifiers may differ in scope from local to global. For instance, while the ORCID has a global scope for researchers, SSN is a code that is associated to citizens having scope limited to a certain country; clearly, both DOI and ISBN have global scope. In an open world in which the data we have about a certain entity can be locally incomplete, it is appropriate to enforce that most of the properties are not mandatory, including identifiers. Thus, not all identifiers might be available at the same time for all entities. Consistency of transactions should be supported based on those identifiers whenever available.

Analogously, we suggest that *reference* should be supported by the Knowledge HUB via the definition of a primary identifier that is used as target of the relations to link entities to one another. Cataloguers may decide to use names (like in libraries), URIs (like in the Semantic Web) or any other identifier (like foreign keys in databases) for reference. In this case, they need to be defined as identifiers in the data model.

Given the focus of libraries so far, *name authority* is traditionally bound to author names and document titles [6]. It is clear that *name authority* in the Knowledge HUB needs be extended to the names of entities of any kind. As described in the previous section, we believe name authority should be supported by means of dedicated data structures that allow names to be described uniformly. They need to accommodate for name variants and variations in multiple languages. The data structures may differ according to the entity type.

Controlled access points play a crucial role during the data integration in the Knowledge HUB. In fact, entities extracted from the legacy datasets may have different identifiers associated to them. When same identifiers are used across datasets (they can be local to a university, local to a country, or global), different mentions in different datasets can be identified and merged quite easily. Notice however that, analogously to databases, identity management should be supported by the Knowledge HUB, and not by the ETL facilities. This is fundamental to guarantee consistency of transactions. Identity management should detect duplicates and merge them such that one single copy of each entity is maintained and retrieved when searching for it.

The more identifiers we define the more effective their identification across information silos will be. Conversely, the absence of common identifiers requires heuristics to be put in place. For instance, identification may fail when a legacy dataset completely lacks of identity management or name authority mechanisms. For example, the institutional repository of the University of Trento does not provide any mechanism to disambiguate locations (of conferences and of editors). Examples of useful heuristics, though restricted to author names only, can be found in [12].

7. Vocabulary Control

Vocabulary control in libraries enforces unique headers for subjects [6], thus making sure there is an unambiguous way to refer to each subject. To this purpose, thesauri can be used, as described for instance

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

in the ISO 25964-1 standard [11]. In a thesaurus, terms denoting subjects are associated to concepts, each of them denoted by a unique identifier, arranged hierarchically from more general (broader) to more specific (narrower) concepts. It is possible to associate multiple synonymous terms to the same concept, in multiple languages.

To make sure that the whole terminology used to describe entities is controlled, we need to develop thesauri not only for subjects, but also to control the terminology used for:

1. the names of the *classes*
2. the names of the *relations*
3. the names and values of the *attributes* (including subject), whenever possible

To deal with controlled values we introduced the data type Concept in the data model in Figure 2. The extended vocabulary control ensures uniformity in the way in which entities are encoded and queried.

To develop vocabularies able to control all the necessary terminology above, in the past years we expressly devised the DERA methodology [8]. The methodology consists of a series of steps and guiding principles in turn inspired to the analytico-synthetic approach created by Ranganathan [10]. The analytico-synthetic approach is at the basis of the Colon Classification and it is known to guarantee the development of high quality vocabularies. DERA supports the development of thesauri compliant with the ISO 25964-1 standard. The interested reader may refer to [23] for additional details about the methodology and its principles. In fact, DERA has been designed to arrange concept hierarchies into three categories, corresponding to the three kinds of terms given above:

1. ***Each entity type generates a hierarchy of category ENTITY*** (the E of DERA) that captures the possible values of the Class attribute (e.g. paper, journal).
2. ***Each relation generates a hierarchy of category RELATION*** (the R of DERA) specifying relation names from more general (e.g. creator) to more specific (e.g., author, painter).
3. ***Each attribute generates a hierarchy of category ATTRIBUTE*** (the A of DERA) that includes the attribute names (e.g., status of a project) and the possible values (e.g., funded, rejected).

For instance, the vocabulary given in

Figure 3 can be associated to the data model in Figure 2. The first term associated to each concept is the preferred term; the other terms are the synonyms. The number represents the concept identifier.

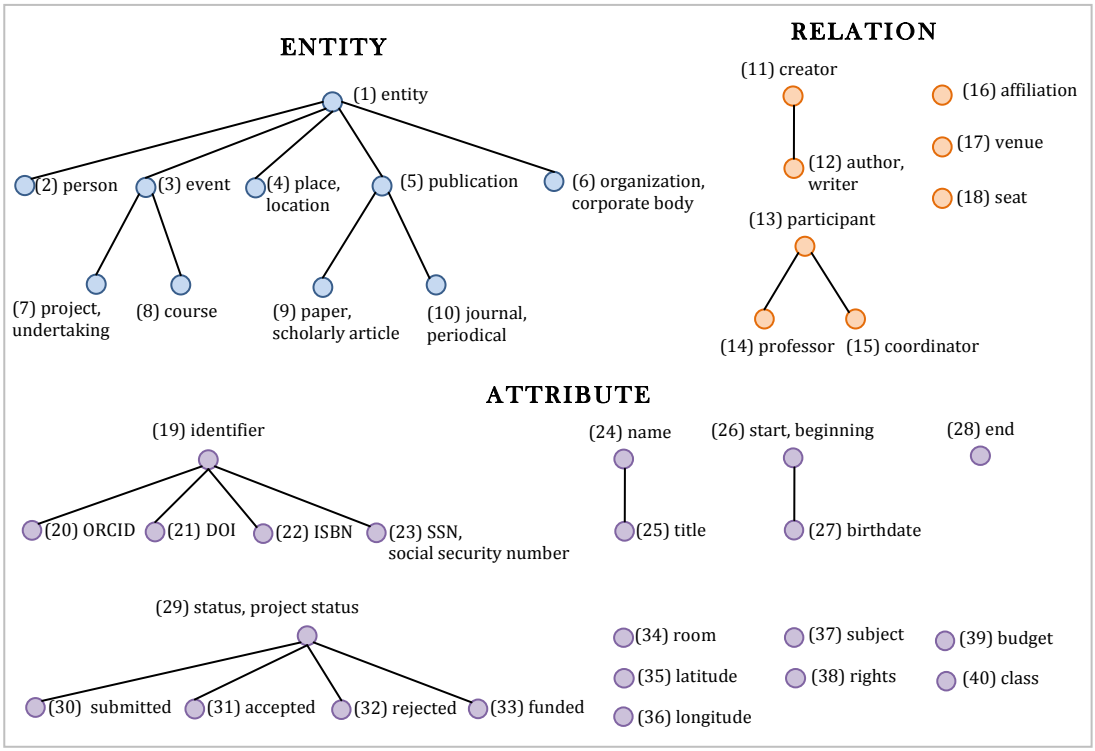

Figure 3 – Example of vocabulary for the data model given in Figure 2

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

8. Data hunting

Suppose that the University of Trento is constituted by two departments, the department of Economics and the department of International Studies. Suppose that the two departments maintain their data in different information silos, as depicted in Figure 4 for the department of Economics and in Figure 5 for the department of International Studies.

DB1-E: Employee

ID	Name	Birthdate	Salary
11	John Doe	1974-09-26	2500
22	Paul Connor	1969-05-12	3000

DB2-E: Publication

ID	Title	Subject	Author	Kind	ISBN	DOI	Rights
1	Markets	Globalization	J. Doe (345); Connor (567)	Paper	*null*	123	CC0
2	Finance	Recession	J. Doe(345)	Journal	456	*null*	BY

DB3-E: Course

ID	Name	From	To	Room	Professor ID	Professor
A01	Economics	2013-09-09	2013-12-20	A203	11	John Doe
A02	International Finance	2013-10-09	2013-12-22	B107	22	Paul Connor

DB4-E: Project

ID	Name	Subject	From	To	Cost	Place	Status
33	Business Models	Business	2010-01-01	2014-12-31	2.5	Trent	proposal
34	Avoiding recession	Recession	2013-01-01	2014-12-31	1.8	Trent	not approved
35	Banks and Finance	Finance	2014-05-01	2015-05-30	0.9	Rome	approved

DB4-E: Participant

ID	Person	Role
33	J. Doe	Coordinator
34	P. Connor	Coordinator
35	P. Connor	Researcher
35	J. Doe	Coordinator

Figure 4 – Example of datasets of the department of Economics

Entities managed include people (e.g., *John Doe*), publications (e.g., *Markets*), courses (e.g., *International Finance*), and projects (e.g., *Banks and Finance*). In particular, they use different HR management systems (DB1-E and DB1-IS) and institutional repositories (DB2-E and DB2-IS), while they use the same systems to manage courses (DB3-E and DB3-IS) and research projects (DB4-E and DB4-IS), but instantiated on different servers and maintained separately. Due to the different users and conventions, even when the systems are the same, the two departments use different terminology to express similar concepts. For instance, the department of Economics uses the term *proposal* to denote a *submitted* project, i.e. the term used by the department of International Studies; similarly, the term *paper* used by the first department corresponds to the term *scholarly article* used by the second department.

DB1-IS: Affiliate

SSN	Name	Birthdate
A31356	Anthony Black	1974-09-26
B25555	Paul Connor	1969-05-12

DB2-IS: Repository

ID	Title	Topic	Author	Type
1	Immigration in Europe	Immigration	A. Black (A31356); Connor Paul (B25555)	Scholarly article
2	The European Markets	Financial crisis	C. P. (B25555)	Journal

DB3-IS: Course

ID	Name	From	To	Room	Professor ID	Professor
B01	Sociology	2013-09-09	2013-12-20	C101	A31356	Anthony Black
B02	Management	2013-10-09	2013-12-22	D202	B25555	Paul Connor

DB4-IS: Project

ID	Name	Subject	From	To	Cost	Place	Status
33	Migrations in Europe	Migrations	2011-01-01	2013-12-31	1.5	Trent	submitted
34	European Finance	Finance	2012-01-01	2014-12-31	2.2	Trent	accepted

DB4-IS: Participant

ID	Person	Role
33	A. Black	Coordinator
34	A. Black	Coordinator
34	Paul Connor	Researcher

Figure 5 – Example of datasets of the department of International Studies

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

9. Populating the Knowledge HUB

The Knowledge HUB is now populated with the data selected and extracted from the identified data sources. In this section we provide an example to illustrate the difficulties that might be encountered.

The main difficulty consists in the identification of all data about a given entity scattered across datasets, e.g. the person *Paul Connor* affiliated to both departments, despite the data sources do not have a common name authority and identity management strategy. For instance, the discovery mechanisms will have to find out (via heuristics) that *Paul Connor* in DB1-E is the same as *Connor* in DB2-E, *P. Connor* in DB4-E, *Paul Connor* in DB1-IS, and *C. P.* in DB2-IS as well as (via identifiers) the same as *Paul Connor* in DB3-E and DB3-IS. Notice how DB1-E and DB1-IS play the role of authority files for people, but only for some of the datasets, within the departments of Economics and International Studies, respectively.

Another important difficulty stands in mapping all the terms describing the entities in the databases to the controlled vocabulary in

Figure 3. For instance, the attribute names *from* and *to* in DB3-E and DB3-IS are mapped to *start* (concept #26) and *end* (concept #28), respectively. The attribute values *proposal* in DB4-E and *submitted* in DB4-IS are both mapped to *submitted* (concept #30).

In this paper this problem is exemplified, but in the reality the heterogeneity of terms can be extreme. Terms used, both in schema and in data, can be difficult to interpret. For instance, terms might be abbreviated and concatenated in various ways (e.g. “pub_cd” to indicate “publication code”) or can be provided in multiple languages. For instance, terms used for keywords in the main institutional repository of the University of Trento appear mainly in English, Italian, Spanish, French and German. Natural Language Processing (NLP) and schema matching tools [16] can aid people in this task.

Figure 6 shows the authority files of the Knowledge HUB that can be generated as result of the integration of these datasets. The authority files follow the data model given in Figure 2. Names are made uniform by means of name authority rules and act as references between authority files. Terms appearing in the graph are made uniform by aligning them to the controlled vocabulary in

Figure 3. In the figure, they are represented with the term, but they can be stored together with their concept identifier.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

Person							
ID	Class	Name	Subject	ORCID	SSN	Birthdate	Affiliation
E-11	Person	John Doe	Globalization, Recession, Business, Finance			1974-09-26	Economics
E-22	Person	Paul Connor	Globalization, **Immigration, Financial crisis**		**B25555**	1969-05-12	Economics, **International Studies**
	Person	**Anthony Black**	**Migrations, Finance**		**A31356**		**International Studies**

Paper							
ID	Class	Title	Subject	DOI	ISBN	Rights	Author
E-1	Paper	Markets	Globalization	123		CC0	John Doe, Paul Connor
E-2	Journal	Finance	Recession		456	BY	John Doe
IS-1	**Paper**	**Immigration in Europe**	**Immigration**				**Anthony Black, Paul Connor**
IS-2	**Journal**	**The European Markets**	**Financial crisis**				**Paul Connor**

Course							
ID	Class	Name	Subject	Start	End	Room	Instructor
E-A01	Course	Economics		2013-09-09	2013-12-20	A203	John Doe
E-A02	Course	International Finance		2013-10-09	2013-12-22	B107	Paul Connor
IS-B01	**Course**	**Sociology**		**2013-09-09**	**2013-12-20**	**C101**	**Anthony Black**
IS-B02	**Course**	**Management**		**2013-10-09**	**2013-12-22**	**D202**	**Paul Connor**

Project								
ID	Class	Name	Subject	Start	End	Budget	Status	Coordinator
E-33	Project	Business Models	Business	2010-01-01	2014-12-31	2.5	Submitted	John Doe
E-34	Project	Avoiding recession	Recession	2013-01-01	2014-12-31	1.8	Accepted	Paul Connor
E-35	Project	Banks and Finance	Finance	2014-05-01	2015-05-30	0.9	Rejected	John Doe
IS-33	**Project**	**Migrations in Europe**	**Migrations**	**2011-01-01**	**2013-12-31**	**1.5**	**Submitted**	**Anthony Black**
IS-34	**Project**	**European Finance**	**Finance**	**2012-01-01**	**2014-12-31**	**2.2**	**Accepted**	**Anthony Black**

Figure 6 – The authority files in the Knowledge HUB populated with the data

It is important to underline the following aspects:

- ***Not necessarily the whole information content that is available in the original datasets needs to be represented in the HUB.*** For instance, salary information that is important to manage employees in DB1-E may be irrelevant for the envisioned centralized services of the Knowledge HUB or may be excluded a-priori because of privacy concerns.
- ***Not necessarily the original datasets can provide all the necessary information.*** For instance, subject information is missing for courses both in DB3-E and DB3-IS.
- ***Some information that is implicit in the original datasets can be reconstructed in the HUB.*** For instance, implicit assumptions that have been made include the fact that all employees are people, and that the affiliation of each person to a certain department depends on the system storing data. This is consistent with Artificial Intelligence theories of generality of knowledge [21].
- ***Each piece of information in the Knowledge HUB is accompanied with meta-information*** such as provenance, cataloguers' comments, timestamps, versioning and so on. This is necessary to reconstruct the origin and time of every piece of information. In Figure 6 this is exemplified by representing in plain text data coming from Economics datasets and in bold data coming from the International Studies datasets.
- ***The content of the HUB needs to be constantly kept aligned with the data sources***: the alignment needs to be done by means of dedicated facilities that run periodically, e.g. once a day, and make sure that the necessary transformations are always applied consistently.

Notice that all these services benefit from the ability of the system infrastructure to fix, to a certain extent, quality issues affecting the data sources or at least to uniform the terminology used by them. For instance, it may fix misspellings.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

Consider now the services described in Section 3. We can develop a service able to address the third example of analytics about funded projects given in Table 2. In fact, the Knowledge HUB - despite the initial noise, fragmentation and diversity of the data sources - can now treat projects uniformly. This is actually one of the main benefits of Digital Universities as we conceive them, and justifies the cost of authority and vocabulary control.

10. Trento as Digital University

At the University of Trento, we are moving the first steps towards implementing the vision presented in this paper. In January 2015, the University launched the *Digital University* initiative that aims to gradually put in place the system infrastructure required. The Knowledge HUB of the University of Trento will provide centralized access to information that is otherwise spread across multiple information silos. In this way, it will be able to offer a broad range of on-line centralized services to a variety of users including students, researchers, professors, and members of the governing body.

The Knowledge HUB is an instance of the SWEB semantic technology developed at the University of Trento in many years of research in Artificial Intelligence, Semantic Web and Digital Libraries (see for instance [17][18][8]). SWEB internally represents data as knowledge graphs. It offers APIs and graphical user interfaces supporting all the modelling and authority control tasks (identity management, name authority and vocabulary control) which are necessary to govern the knowledge graph. These facilities are integral part of what we call the *Knowledge Operating System* (KOS).

Figure 7 provides a screenshot of the user interface that allows the definition and visualization of the data model.

Figure 8 provides a screenshot of the user interface that allows to search within the controlled vocabulary; when a concept is clicked it is possible to navigate the network of semantic relations between concepts.

The system infrastructure is completed with ETL facilities supporting the selection and interpretation of data from the institutional legacy datasets and the creation of the knowledge graph in adherence with the data model and the authority control rules. Data about the same entity scattered across different datasets is correlated by the Knowledge HUB by means of identifiers or heuristics. Duplicates are detected and merged into a single entity. Currently the knowledge graph of the University of Trento contains about 1 million entities. Entities primarily include people, organizations, locations, publications of various kinds, dissertations, patents, courses, and projects.

In addition, the Knowledge HUB offers APIs exposing the basic search facilities necessary for the development of the services. Services being developed include an institutional portal offering a unified view of the faculty members and departments in multiple languages, and an institutional dashboard providing analytics about the quality of research conducted by the faculty members. In particular, the institutional portal will leverage on the capacity of the Knowledge HUB to integrate and correlate data about the faculty members across the original information systems and to codify information using the concepts defined in the vocabularies such that they can be searched and visualized in multiple languages.

Legal challenges we need to face include the modalities by which we ensure the privacy of the users and how we comply with IPR. This is tackled by ensuring governance and privacy-by-design principles [22]. In practice this means that privacy has to be considered a fundamental functional requirement since the design of the entire system infrastructure. Understanding what kind of privacy concerns and principles we need to consider is part of this process which is informed by the legal office of the University. In particular, the Knowledge HUB needs to maintain the data in compliance with the national laws, and guarantee secure access to only authorized users. In terms of IPR, we decided to promote and support the download of the Open Access publications through our institutional websites.

Organizational challenges [20] are tackled by employing an interdisciplinary pool of people skilled in ICT and Library & Information Science that closely collaborate with representatives from the various departments. In particular, we created dedicated boards responsible for the definition and maintenance of an official subject vocabulary per academic department, handling domain-specific subjects. There is a representative professor for each department who coordinates the collection of the ground topics that are mapped with standard thesauri and arranged into hierarchies by the experts. The experts are responsible of the metadata quality, the entity cataloging, and of the correct interpretation of data to support institutional decision-making.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

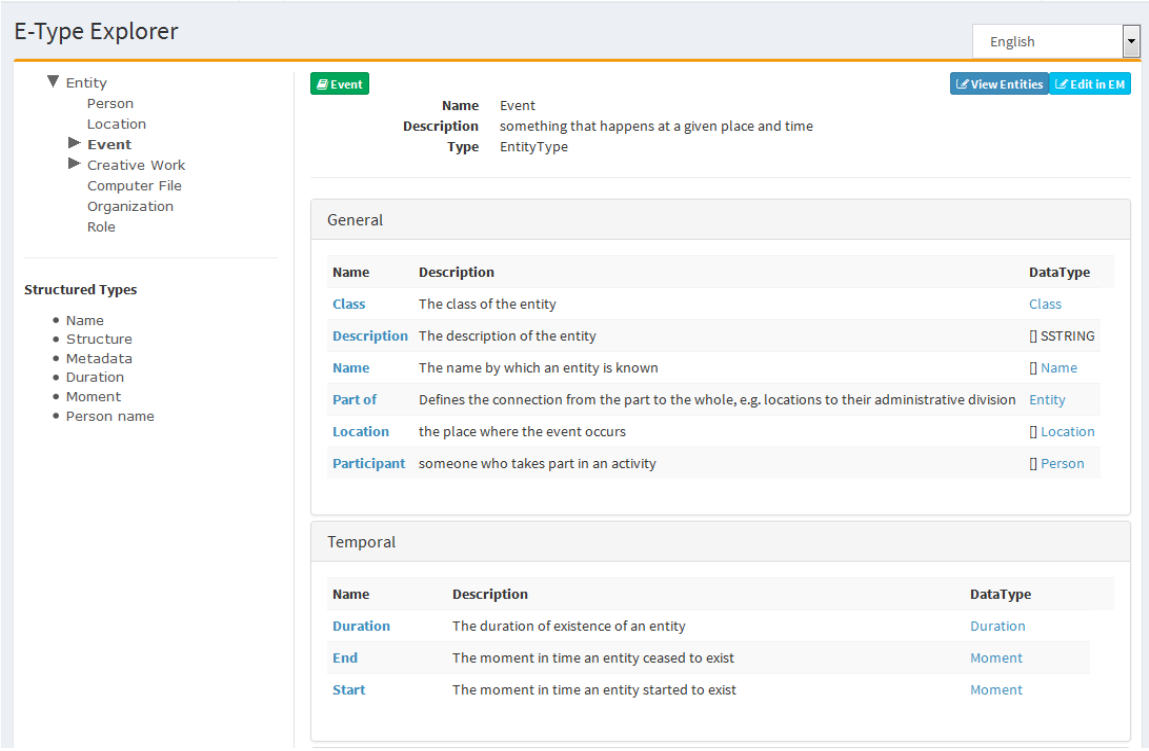

Figure 7 – A screenshot of the Data Modelling facility of the Knowledge HUB at Trento

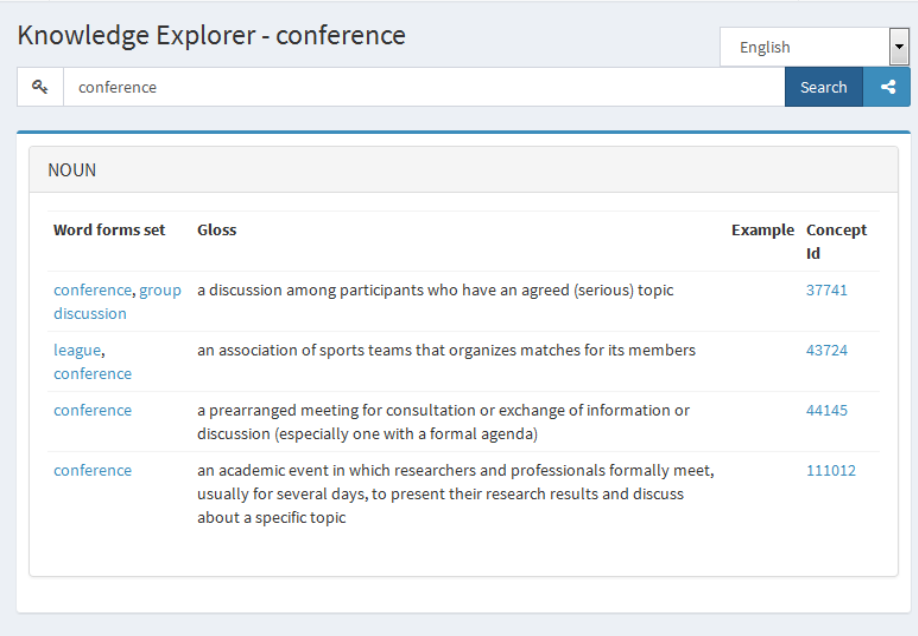

Figure 8 – A screenshot of the Knowledge Explorer facility of the Knowledge HUB at Trento

The University of Trento expects that the initiative will contribute to the establishment of a new data curation culture leading to a gradual improvement in the quality of the data managed (for instance by informing the people responsible of the source datasets about how to improve their data quality), in the efficiency in their treatment, and in the capacity to support self-assessment and institutional decision-making.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Quarterly** December 2016 - Vincenzo Maltese - http://dx.doi.org/10.1080/01639374.2016.1245231. **Cataloging & Classification Quarterly** is available online at: http://www.tandfonline.com/

11. Conclusions

In this paper, we introduced *Digital Universities* that we define as *a set of key resources and tools appropriately organized to effectively support universities' users*. We described how the new concept emerges as natural extensions of Digital Libraries. We described the key knowledge assets to be managed and the services required. The envisioned system infrastructure is centered on the notion of Knowledge HUB, a platform that by means of authority control is able to address the data fragmentation and data diversity that is intrinsic in universities. Such mechanisms support the uniform cataloguing of all the key knowledge assets of the university such that services can leverage on high quality data. We illustrated how the authority files contained in the HUB are populated and the tasks that need to be supported in order to provide centralized access to information that is otherwise spread across multiple information silos. In this respect, the HUB acts as a trusted proxy between the services and the original data sources. The APIs provided by the HUB can be exploited by universities to develop innovative centralized services to their users.

We are aware that this work opens new challenges to be address in the following years. Challenges includes the definition of a core standard data model for universities, the identification of the most appropriate methodologies and data representation paradigms to be adopted, as well as the development of adequate system infrastructures and tools. We already started discussing these themes in informal workshops with colleagues working in other universities in Europe, China, India and Mongolia.

Acknowledgements

In memory of Umberto Maltese (father of Vincenzo Maltese) who died in July 2016. We would like to thank all the people who have been collaborating with us to the Digital University project in Trento during the first two years of activity. Special thanks to Stella Margonar who heavily contributed to the development of the technologies at the basis of the Knowledge HUB, to the former general director of the University Giancarla Masè and the IT director Andrea Mongera who strongly supported the initiative since its first stirrings. We are grateful to our rector Paolo Collini, the current general director Alex Pellacani, the ICT rector's delegate Renato Lo Cigno, the head of the library services office Francesca Valentini, the head of the legal office Lucia Anna di Paolo, the head of the research support and knowledge transfer division Vanessa Ravagni, the head of the communication division Paola Fusi and her collaborators Laura Salvetti and Miriam Sebastiani, the head of the information systems office Mauro Filippi, the professors who are developing with us the subject classifications of the various academic departments, and the dozens of other people of various offices and departments of the University of Trento who have been collaborated with us. We also want to thank Carol Ellerbeck for the interesting and lovely discussions we had together at the Harvard Business School in US. This research has been partially funded by the European Community's 7th Framework Program "Smart Society", under grant agreement n. 600854 (http://www.smart-society-project.eu/).

References

[1] Hannemann, J. and Jürgen, K. 2010. *Linked data for libraries*. World library and information congress of the IFLA.

[2] Martin, K. E., and Mundle, K. 2014. *Positioning Libraries for a New Bibliographic Universe*. Library Resources & Technical Services, 58(4), 233-249.

[3] Byrne, G., and Goddard, L. 2010. *The strongest link: Libraries and linked data*. D-Lib magazine, 16(11), 5.

[4] Singer, R. 2009. *Linked Library Data Now!* Journal of Electronic Resources Librarianship 21 no.2: 114-126.

[5] Saur, K.G. 2009. *Functional Requirements for Authority Data – A Conceptual Model*. IFLA Working Group on Functional Requirements and Numbering of Authority Records. Edited by Glenn E. Patton.

[6] Zeng, M. L., Žumer, and M., Salaba, A. 2011. Functional Requirements for Subject Authority Data (FRSAD): A Conceptual Model (Vol. 43). Walter de Gruyter.

[7] Saur, K.G. 1998. *Functional requirements for bibliographic records: final report*. IFLA Study group on the Functional Requirements for Bibliographic Records.

[8] Giunchiglia, F., Dutta, B., and Maltese, V. 2014. *From Knowledge Organization to Knowledge Representation*. Knowledge Organization, 41(1), 44-56.

[9] Teets, M., and Goldner, M. 2013. *Libraries' role in curating and exposing big data*. Future Internet, 5(3), 429-438.
[10] Ranganathan, S. R. 1977. *Prolegomena to library classification*. Asia Publishing House.
[11] ISO 25964-1:2011. Information and documentation – Thesauri and interoperability with other vocabularies – PART 1: Thesauri for information retrieval.
[12] Fan, X., Wang, J., Pu, X., Zhou, L., and Lv, B. 2011. *On graph-based name disambiguation*. Journal of Data and Information Quality (JDIQ), 2(2), 10.
[13] Kroeger, A. 2013. The road to BIBFRAME: the evolution of the idea of bibliographic transition into a post-MARC future. Cataloging & Classification Quarterly, 51(8), 873-890.
[14] Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society. Edited by D. Miorandi, V. Maltese, M. Rovatsos, A. Nijholt, J. Stewart. Springer, 2015.
[15] Davenport, T. H., and Patil, D. J. 2012. *Data scientist: The sexiest job of the 21st century*. Harvard Business Review, 90, 70-76.
[16] Shvaiko, P., and Euzenat, J. 2005. *A survey of schema-based matching approaches*. Journal on Data Semantics, 4, 146-171.
[17] Giunchiglia, F., Maltese, V., and Dutta, B. 2012. *Domains and context: first steps towards managing diversity in knowledge*. Journal of Web Semantics, 12–13, 53-63.
[18] Giunchiglia, F. 2006. *Managing Diversity in Knowledge*. Invited talk at ECAI.
[19] Maltese, V., and Giunchiglia, F. 2016. *Search and Analytics Challenges in Digital Libraries and Archives*. ACM Journal of Data and Information Quality, Vol. 7, No. 3, Article 10.
[20] Bygstad, B., Ghinea, G., and Klæboe, G. T. (2009). *Organisational challenges of the Semantic Web in digital libraries: A Norwegian case study*. Online Information Review, 33(5), 973-985.
[21] McCarthy, John. *Generality in artificial intelligence*. Communications of the ACM 30.12 (1987): 1030-1035.
[22] Hoepman J. 2014. *Privacy design strategies*. ICT systems security and privacy protection, Springer Berlin Heidelberg, 446-459.
[23] Giunchiglia, F., Dutta, B., Maltese, V., Farazi, F. 2012. *A facet-based methodology for the construction of a large-scale geospatial ontology*. Journal on Data Semantics, 1 (1), pp. 57-73.
[24] Zablith, F., d'Aquin, M., Brown, S., & Green-Hughes, L. 2011. *Consuming linked data within a large educational organization*. Consuming Linked Data Workshop.
[25] Halaç, T. G., Erden, B., Inan, E., Oguz, D., Gocebe, P., & Dikenelli, O. 2013. *Publishing and linking university data considering the dynamism of datasources*. 9th International Conference on Semantic Systems, pp. 140-145, ACM.
[26] Börner, K., Conlon, M., Corson-Rikert, J., & Ding, Y. 2012. *VIVO: A semantic approach to scholarly networking and discovery*. Synthesis lectures on the Semantic Web: theory and technology, 7(1), 1-178.
[27] Astrova, I. (2009). Rules for mapping SQL relational databases to OWL ontologies. In Metadata and Semantics (pp. 415-424). Springer US.
[28] Sicilia, M. A., & Lytras, M. D. (Eds.). (2008). Metadata and semantics. Springer Science & Business Media.