This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

Digital transformation challenges for universities: ensuring information consistency across digital services

Abstract. Universities struggle to offer complete, up-to-date and consistent information about their key assets to their numerous users across various digital services and communication channels. Key assets include people, papers, books, dissertations, patents, courses and research projects. The main difficulty stands in the intrinsic *data fragmentation* and *data diversity*: data about the key assets is scattered across multiple information silos, data is often duplicated and difficult to correlate due to the diversity in format, metadata, conventions and terminology used. We illustrate how this difficulty can be tackled and describe the work carried out at the University of Trento in Italy.

Keywords: Entity-relationships modeling, Controlled vocabularies, Digital repositories, Discovery platforms and services, Institutional repositories, Metadata

1. Introduction

The term *digital transformation* is often used to indicate a set of mainly technological, cultural, organizational, social, creative and managerial changes. Digital transformation goes beyond the simple adoption of new technologies and makes it possible to provide services, supply goods, live experiences, find, process and make accessible large amounts of content regardless of the real availability of resources (human, material, intellectual, economic, etc.), pervasively creating new connections between people, places and things [1]. Even though it brings new opportunities, digital transformation also poses new challenges for Communication and IT departments of universities. In fact, these challenges have been the focus of the 2018 conference organized by EUPRIO, the European Association of Communication Professionals of Higher Education[i].

We concentrate on *digital information challenges* faced by universities nowadays. Universities need to provide detailed information about a variety of key assets to their users. This includes information about their students, employees, professors and researchers, their publications and patents, the courses they teach and the research projects they conduct. It is however difficult for universities to present a complete, up-to-date and

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

consistent view about their key assets across the different digital communication channels and digital services employed.

For example, it may happen that a certain person is an *associate professor* according to the human resources (HR) management system (the main authority for such information), a *research fellow* on the main institutional portal (the portal is outdated), and a *post-doc researcher* on the department website (the website is not only outdated, but it uses different terminology with respect to the institutional portal).

The roots of this difficulty stand in the inherent complexity of the IT university ecosystem. On the one hand, such complexity is unavoidable and actually valuable. In fact, the usage of different IT systems is functional to the myriad of business processes that universities need to conduct: institutional communication, library management, HR management, teaching and student support, research and technology transfer support, project management and fundraising, financial support, IT support, legal support, logistics, strategic planning, and many others. Each IT system targets a specific business process with a controlled number of users, specific key assets and confined responsibility. On the other hand, the *data fragmentation* and the *data diversity* increase with the growth of data and systems employed, thus generating a sort of entropic effect:

- data about the key assets is scattered across multiple information silos;
- data differs in format, metadata, conventions and terminology used;
- data gets duplicated;
- discrepancies and conflicts increase because different versions and descriptions of the same assets coexist in multiple IT systems.

This difficulty is common to many other large-scale organizations. In fact, Gartner says that a significant number of organizations, unable to organize themselves effectively for 2020, will experience an information crisis, due to their inability to effectively value, govern and trust their enterprise information [2].

Maltese and Giunchiglia [3] proposed a general solution to address *data fragmentation* and *data diversity* in universities. By adapting and extending the notion of digital library to universities, they introduced the notion of *digital university*, defined as a set of key resources, methodologies and tools appropriately organized to effectively support universities' users. The proposed solution stands in (a) addressing *data diversity* via the adoption of Library & Information Science (LIS) methodologies and tools to curate *data and metadata quality* and (b) addressing *data fragmentation* via the adoption of *data integration* methodologies and tools.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

In this paper, we provide the description of the infrastructure, the tools and the services that - by following the methodology initially proposed by Maltese and Giunchiglia [3] - we actually put in place at the University of Trento in the context of the Digital University initiative. We also illustrate the main challenges we faced and the lessons learned. The infrastructure that we implemented follows a Hub-and-Spoke paradigm[ii]. The Hub is an IT system that collects data extracted from various information silos and encodes it in a uniform schema, format and terminology. It provides centralized access to a number of spokes, each of them being an IT system developed to support a different digital service. At regular basis, data is selected and directed to each spoke based on what is strictly required by the digital service.

The rest of the paper is organized as follows. In Section 2, we illustrate state of the art and related work. In Section 3, we describe the work done in Trento. In Section 4, we summarize the challenges faced and the lessons learned. Finally, Section 5 concludes the paper.

2. State of the art and related work

Several research communities address data fragmentation and data diversity. In the following, we focus on the solutions proposed by Business Intelligence (BI) and LIS.

The primary purpose of BI is to support decision-making in organizations [4]. Data-driven decision-making refers to the practice of basing decisions on the analysis of data rather than purely on intuition [5]. Therefore, data needs to be appropriately collected and prepared. To this end, *data integration* is a fundamental technique in BI to tackle the initial data fragmentation and diversity. In fact, data integration is a process that combines data from different sources and provides users with a uniform view of data [6]. Two main alternative approaches exist. In *federated systems*, data is logically combined at query time. In *centralized systems*, data is physically combined in a data warehouse via Extract, Transform and Load (ETL) procedures. The Extract phase deals with the selection, assemblage, analysis and processing of data. The Transform phase takes care of converting data into a standard format. The Load phase imports data into the data warehouse. The centralized approach ensures there is one trusted proxy providing data in a timely manner and uniformly. Data warehousing is a fundamental tool of BI, and metadata plays a key role because of the complexity of the data migration process. Therefore, data warehouse teams and business users must understand myriad characteristics of data to manipulate and use it effectively [7].

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

Library Science is traditionally concerned with archiving texts and organizing storage and retrieval systems to give efficient access to texts [8]. LIS is the technical and technological innovation of Library Science that employs information technology for documentation and library services [9]. Libraries have a strong tradition in data and metadata curation, especially in terms of standard data models for the representation of intellectual and artistic creations [10]. Metadata about them includes title, subject, and authors. Authority control makes sure that each entity is assigned a unique header such that each entity can be uniquely identified and referred to [10]. Unique headers include names and identifiers. Similarly, vocabulary control enforces the usage of standard terms to unambiguously refer to each subject [11]. In controlled vocabularies, standard terms are arranged hierarchically from broader to narrower terms [12]. For instance, in biology we may establish that the standard term to denote "any malignant growth or tumor caused by abnormal and uncontrolled cell division" is *cancer*, that cancer is a *disease* (broader term) and that *melanoma* is a type of cancer (narrower term). Thus, a user searching for cancer can be directed also to texts about melanoma. Altogether, the adoption of these practices allows controlling diversity and obtaining *high quality data* that in turn ensures high precision and recall in search. Data fragmentation is addressed in libraries by employing standard data exchange protocols; see for instance the OAI-PMH framework [13].

A few initiatives recently provided solutions to support storing, searching, browsing, visualizing and sharing scholarly data. *Linked Universities*[iii] and *VIVO*[iv] [14] rely on Semantic Web technologies to represent and store data in the RDF standard model[v] and retrieve it using the SPARQL query language[vi]. However, it has been already observed that these initiatives offer limited support to tackle data diversity and data fragmentation [3]. In fact, even though URIs play the role of unique headers, nothing prevents the usage of different URIs for the same entity across datasets. Duplicates are handled at importing time by discovering and linking them automatically. The discovery of duplicates can be achieved for instance by means of String similarity. The linking is typically done by defining the owl:sameAs relations between entities, i.e. the property that - by linking two individuals - specifies that they are actually the same[vii]. This means that duplicates remain unmerged. As a result, queries may return multiple equivalent entities that need to be reconciled before exploiting and visualizing the results. Even though URIs can be used to link terms to external vocabularies, these approaches do not seem to provide any facility or suggest any methodology to effectively control and enforce terminology.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

3. The Digital University solution in Trento

The purpose of the Digital University initiative, launched in 2015 and still running at the University of Trento in Italy[viii], is to turn data into a valuable asset through governance strategies that ensure quality and facilitate the re-use of available information. The solution we adopted follows the methodology described in [3] and is based on the notion of *digital university*. We understand digital universities as natural extensions of digital libraries with a broader scope in entity types and services. Data fragmentation and data diversity are tackled by adopting a combination of both LIS and BI approaches. Data modeling, authority and vocabulary control play a fundamental role in data curation. With the Transform phase of the ETL process, *data diversity* is addressed by codifying data uniformly in schema and terminology, and by consistently assigning a unique identifier to data about the same entity initially scattered across different information silos. With the Load phase, *data fragmentation* is addressed by collecting and pulling together into the data warehouse data about the same entity. In the following, we describe the system infrastructure and the steps of the methodology.

The system infrastructure

The system infrastructure that we implemented (Figure 1) follows a centralized data integration approach based on the Hub-and-Spoke paradigm. The Hub is a data warehouse that receives data from the ETL facilities and, in turn, it provides centralized access to a number of spokes. Each spoke is an IT system that supports a different digital service. We envision different categories of services. *Discovery services* provide basic browsing and search functionalities. *Communication services* convey institutional information to university stakeholders. *Interoperability services* support data exchange; in particular *Open Data services* offer public re-usable data [15]. *Predictive & data analytics services* support institutions in decision-making processes [16].

Hub-and-Spoke architectures give concrete advantages to organizations in maximizing the value of their data [17]. They represent a more efficient and scalable alternative to point-to-point communication in that the number of connectors between systems is reduced drastically, thus reducing complexity and maintenance costs. In fact, a Hub-and-Spoke architecture with n information silos and m services requires only n + m instead of the n x m connectors required by point-to-point communication.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

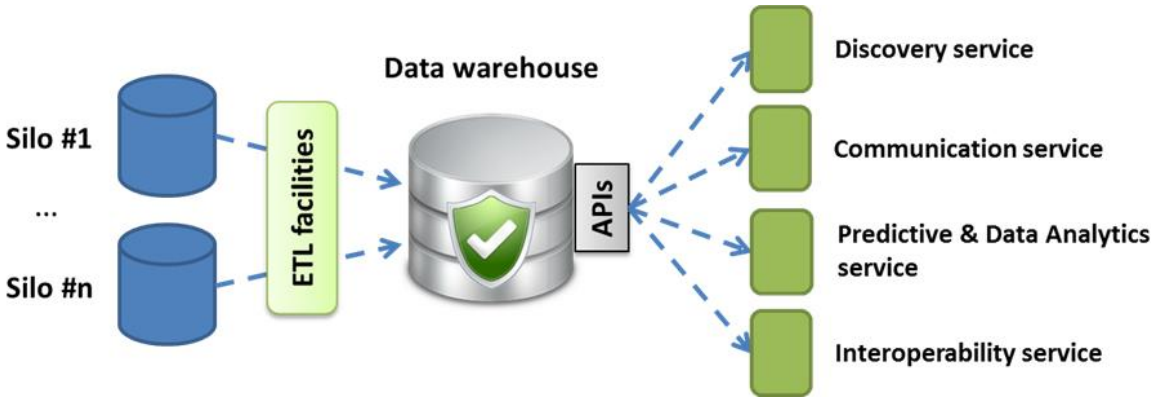

Figure 1 - The system infrastructure of Digital Universities

Step 1: Collecting service requirements

The first step of the methodology consists in collecting the requirements of the digital services to be developed. We targeted two initial services: (1) an institutional portal able to provide an overall view of the university with its institutional and organizational units, as well as its affiliates with a webpage for each of them providing in one single place their main teaching and research assets; (2) an institutional dashboard able to offer useful data analytics. To collect requirements, we interviewed potential users of both services. For the portal, we interviewed students and technical, administrative and academic staff. For the dashboard, we interviewed the heads of our academic departments and the director general. We also reviewed the portals of the top 30 universities according to The Times Higher Education. Our colleagues of the Communication division designed the mockups of the services. We also conducted a user study to validate and refine the mockups.

Step 2: Setting up the data model

The data model serves the dual purpose to drive the ETL process and to set up the hosting schema of the Hub. We developed the data model by extending a model previously employed in similar projects we carried out in collaboration with the Knowdive research group[ix]. The data model accounts for all the entity types and properties necessary for the targeted digital services and includes people, organizations, places, files, papers, books, patents, dissertations, research projects and courses. We inferred what is needed from the mockups of the digital services. This is consistent with modeling methodologies where mockups play the role of competency questions [18, 19].

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

Step 3: Setting up authority control

We se tup entity identifiers and policies governing the form and selection of entity names. To establish identifiers, for each entity type we recognized those properties that can be used to uniquely identify data about the same entity across different data sources. For instance, at the University of Trento both people and organizational units are assigned a unique identifier that is used consistently in the most important IT systems employed. For projects, we used a combination of their name, the starting year and the funding call. To make sure each entity is assigned exactly one name, we designated authorities among the University archives from where we can extract the official name of each entity. Multiple identifiers can be defined for each entity type.

Step 4: Setting up the controlled vocabularies

We defined a controlled vocabulary that includes the standard term, in English and Italian, for each of the entity properties and corresponding allowed values. This is similar to what is typically done for subjects in LIS. See the examples in Table 1 and

Standard Term	**Definition**
abstract	a sketchy summary of the main points of an argument or theory
preprint	a digital draft of a scholarly article before being peer-reviewed
postprint	a digital draft of a scholarly article after it has been peer-reviewed
editorial version	the final version of a peer-reviewed scholarly article in the publishing layout

Table 2. They provide the standard terms for the allowed values of the property "status" of the entity type "project", and of the property "version" of the entity type "file", respectively.

Standard Term	**Definition**
submitted	referred for judgment or consideration
reviewed	appraised critically
pending	awaiting conclusion or confirmation
rejected	something or someone judged unacceptable
funded	furnished with funds

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

unfunded	not furnished with funds
accomplished	successfully completed or brought to an end

Table 1 – standard terms for the values of the property "status" of the entity type "project"

Standard Term	**Definition**
abstract	a sketchy summary of the main points of an argument or theory
preprint	a digital draft of a scholarly article before being peer-reviewed
postprint	a digital draft of a scholarly article after it has been peer-reviewed
editorial version	the final version of a peer-reviewed scholarly article in the publishing layout

Table 2 – standard terms for the values of the property "version" of the entity type "file"

Particular effort was required to define the standard terms for types of positions occupied by our affiliates and for types of organizations. We reviewed the full list of types appearing in University archives. We found around 250 different position types and around 100 organization types. This very fine-grained classification, necessary for administrative purposes, is not appropriate for communication purposes. Therefore, we grouped together similar types and defined a standard term for each group. We obtained 42 groups for positions and 14 groups for organizations. Examples are reported in Table 3 and

Original Organization Types	**Standard Term**	**Definition**
Laboratory	Unit	an organization regarded as part of a larger social group
Working Group		
Institutional or organizational unit		
Project or Service group		
Generic Organizational Unit		
Research group		

Table 4.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

This mapping between University types and standard terms can be configured at any time by University staff by means of a system we expressly developed. We also establish the relative importance of positions and units among them by associating a different weight to each type. For instance, the weight currently associated to "rector", "full professor" and "associate professor" are 100, 450 and 470, respectively. For instance, this allows us to order positions when they appear together for the same person on the portal.

Original Position Types	**Standard Term**	**Definition**
Research fellow	Research fellow	a researcher hired for a limited time period
Researcher within a research project		
Fixed term researcher of type A		
Fixed term researcher of type B		
External researcher		
Researcher with double affiliation		

Table 3 – an example of standard term for position types

Original Organization Types	**Standard Term**	**Definition**
Laboratory	Unit	an organization regarded as part of a larger social group
Working Group		
Institutional or organizational unit		
Project or Service group		
Generic Organizational Unit		
Research group		

Table 4 – an example of standard term for organization types

Step 5: Data hunting

We assessed our university IT systems in order to identify possible sources for the data necessary to fill the data model designed to support the digital services. We identified overall five useful data sources. Yet, given that these sources do not provide all the data required by the digital services, we had to develop two additional

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

archives. The first can be seen as an authority file as it was developed to be able to associate names and descriptions in English and Italian to institutional and organizational units. In fact, available datasets only provide their names in Italian. The second was developed to allow people to provide their personal photos, CVs, notices, office hours and thesis proposals.

Step 6: Developing the ETL facilities

This is the most expensive step of the overall methodology. In fact, significant time and effort is required to understand the schema, terminology and conventions used by each of the data sources. The Extract procedures select relevant data from the data sources and encode it in a standard format. We decided to adopt JSON[x] as it allows the encoding of structured information. The technicians of the IT division gave us access to *database views* expressly arranged to provide access to relevant data only. Among other things, this allows us to comply with privacy and security requirements and makes system maintenance easier. In fact, such database views can be seen as *contracts* that cannot be violated even in case the data source changes, e.g. because of an update of the corresponding IT system. The Transform procedures convert JSON data in compliance with the data model and the controlled vocabulary previously prepared:

- property names are mapped to the corresponding standard name in the controlled vocabulary;
- values are formatted according to the expected data type;
- textual values are mapped to the corresponding standard terms in the controlled vocabulary;
- each piece of data is extended with the unique identifier of the corresponding entity, such that the Hub is able to recognize and merge them appropriately.

Finally, the Load procedures ensure JSON data is loaded to the Hub. The ETL facilities run overnight to make sure the Hub and the spokes receive up-to-date data every day.

Step 7: Implementing the services

Below we provide a screenshot of the two services we developed. The corresponding spoke receives data from the Hub via RESTful APIs that represents the standard programming solution to develop Web Services[xi]. To ensure efficiency at query time and to comply with privacy regulations, data required by each service is stored in distinct ElasticSearch indexes locally to each spoke. ElasticSearch is free and constitutes a very effective solution to store and retrieve data[xii]. In fact, it is known to be extremely fast, scalable, flexible and reliable.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

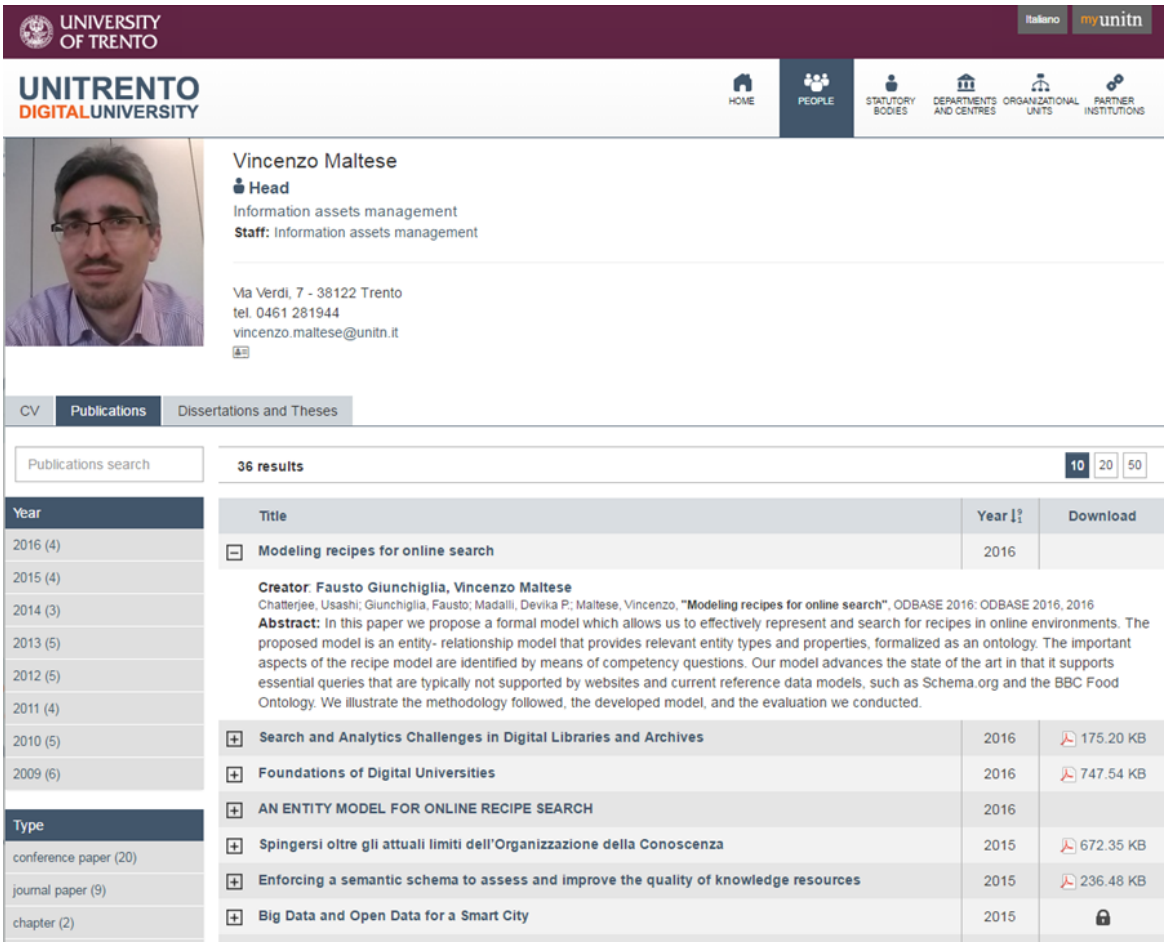

Figure **2** shows a page of the portal. For each person, it offers position and contact information as well as personal teaching and research assets. For the person in the picture, only publications and dissertations are available. Whenever available, the portal also provides information about courses and projects. Users can browse information by selecting values from faceted filters that appear on the left side of the page. For instance, users can filter publications by year of publication, type and co-author. When users select a certain publication, they can visualize corresponding metadata. Users can freely download publications in Open Access; see the download column in

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

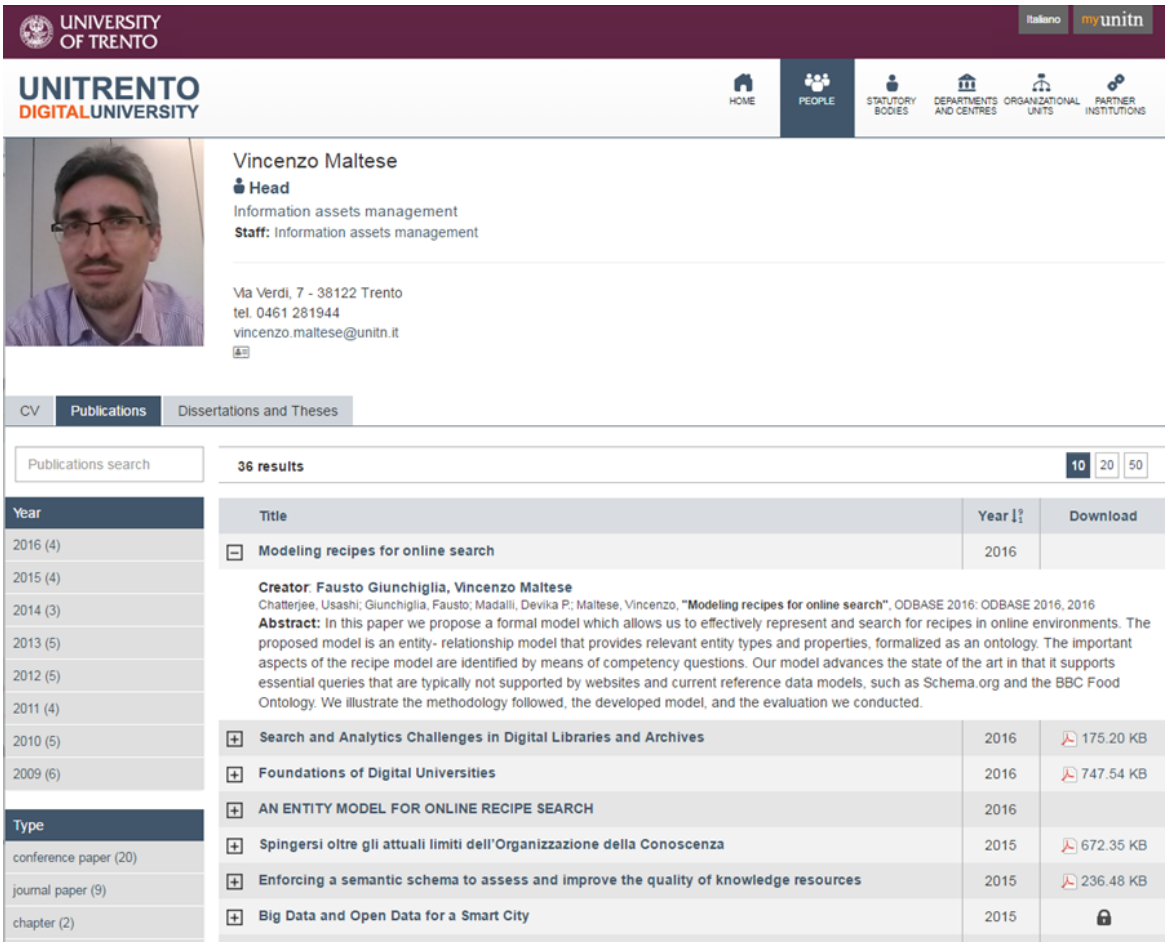

Figure **2**. Currently, more than 2500 unique users visit the portal every day.

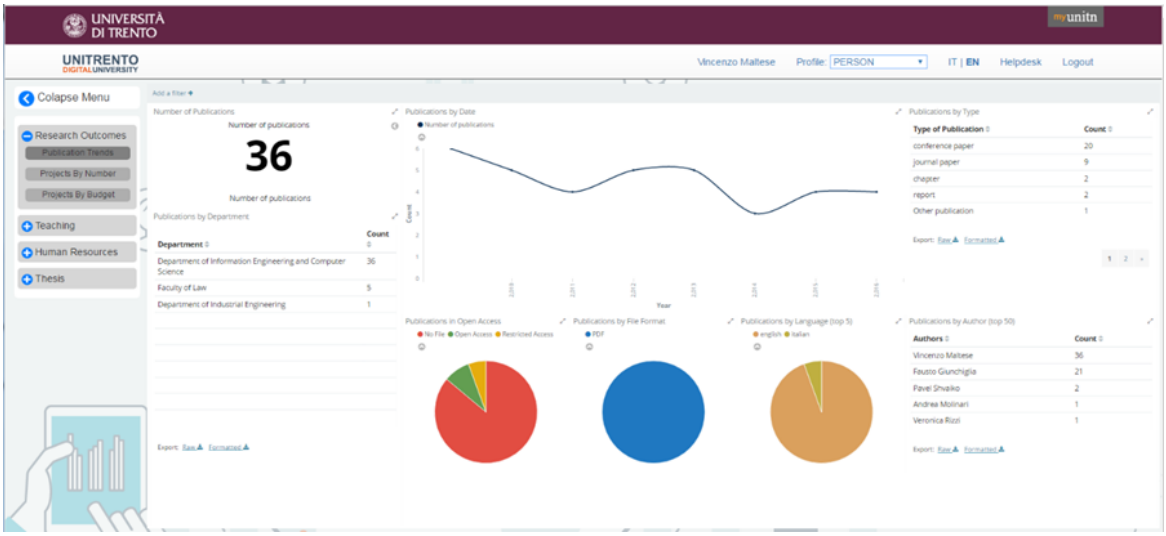

Figure **3** shows a screenshot of the dashboard. The selected page provides aggregated information about the publications of the person in

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

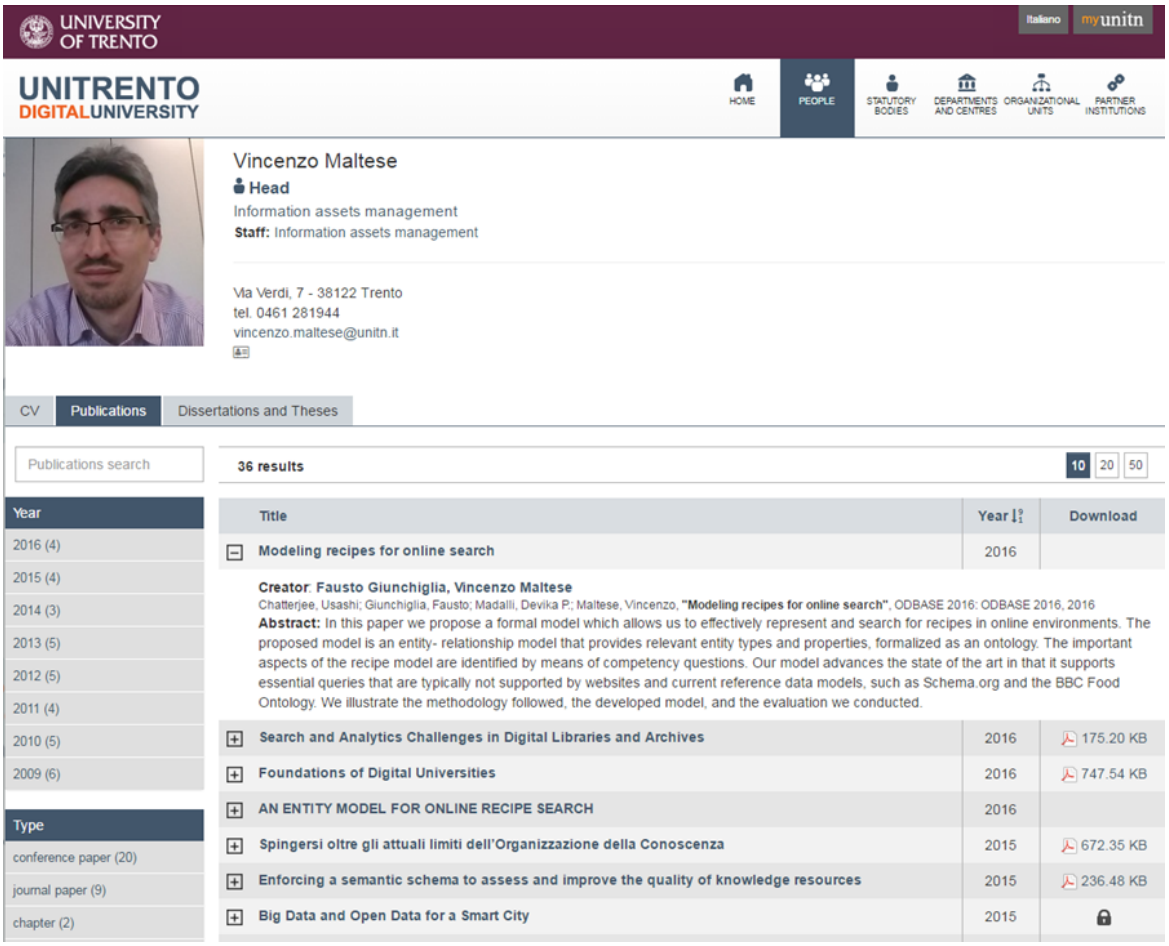

Figure **2**. As you can notice, data is aligned and consistent between the two services, both in terms of figures and terminology used. Heads of academic departments can visualize analytics about the whole department, while other affiliates can only visualize analytics about themselves.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

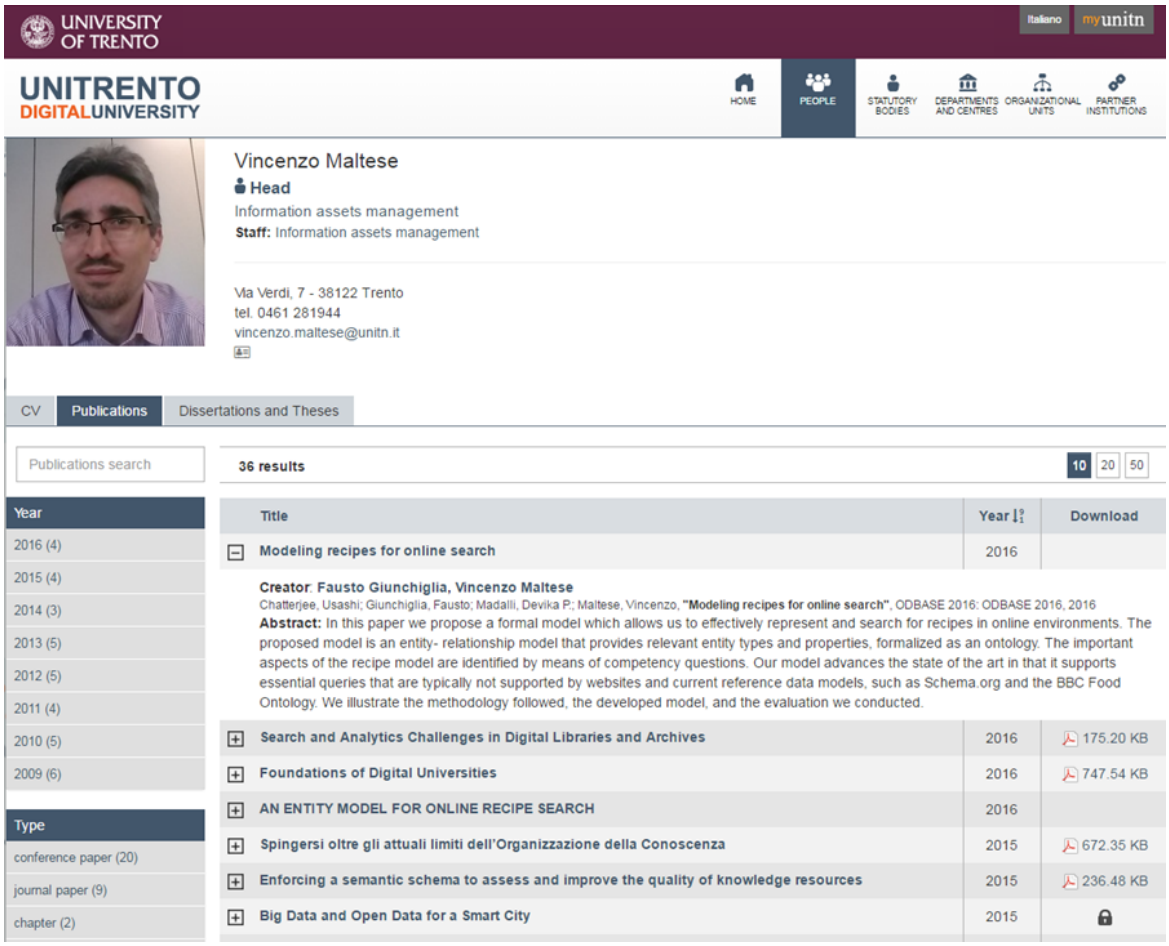

Figure 2 - The Digital University portal, available at the following address: www.unitn.it/du/en

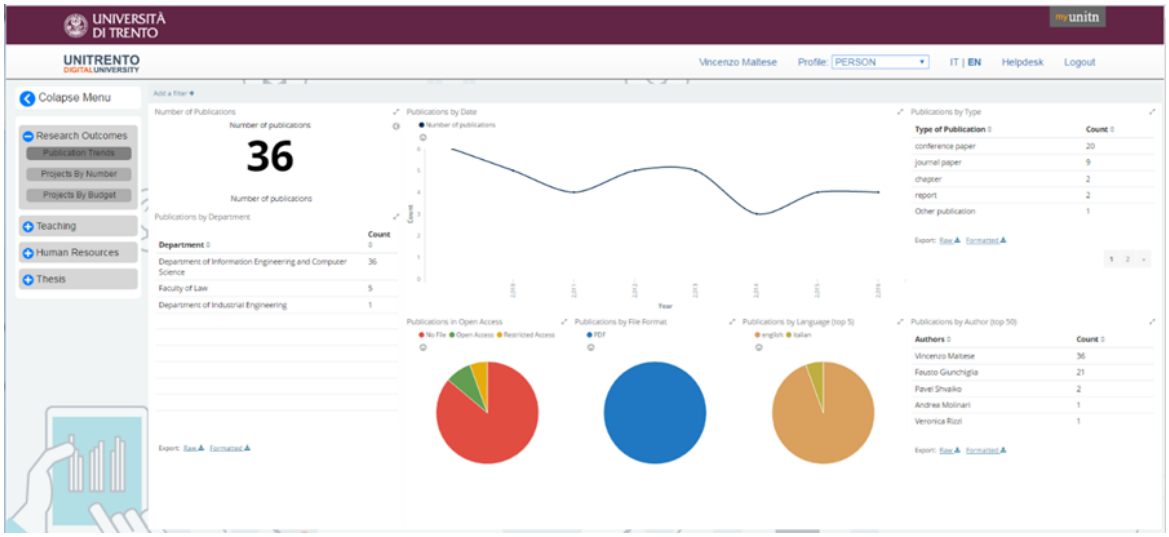

Figure 3 – The University dashboard: publication trends of a person (for authorized users only)

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

4. Challenges faced and lesson learned

Maltese and Giunchiglia [20] briefly described challenges that people typically faced when dealing with advanced search and analytics in digital libraries and archives. Here, we provide a description of the challenges we actually faced so far and the solutions we employed in Trento.

Organizational challenges

Organizational challenges pertain to the obstacles that need to be overcome to move from consolidated practices and standards to new ones, the difficulty to allocate and manage necessary resources, and to identify people with the required skills.

We had to convince the university governance that developing a new IT infrastructure and new digital services is worth not only the investment, but also a pressing need. We met this goal by providing concrete examples of problems that need to be solved and by showcasing what we can achieve. In particular, the Knowdive research group conducted a small-scale pilot project in 2015 that lasted one year and showed that by integrating data extracted from available datasets we can actually reconstruct a comprehensive and coherent picture of what is relevant to each university affiliate. We presented our results at the Academic Senate and the Council of Directors.

We also had to understand the core skills required to carry out the entire project. We formed an interdisciplinary team of people skilled in Information & Communication Technology (ICT) and LIS that closely collaborated with the Legal, the IT, the Library and the Communication departments of the University. We believe that the achievements we have been able to accomplish would not have been possible without such tight collaboration. We set up a clear project plan, with defined tasks, deadlines and responsibilities.

Technical challenges

Technical challenges include the difficulties concerned with the identification of appropriate tools and supporting technologies. The experience we gained tells us that technology selection as well as implementation choices should be made by ensuring that the system satisfies the following fundamental requirements:

- **Maintainability and scalability**: the system should be designed to facilitate future extensions and it should be able to scale with the increase in the amount of data sources and services. As stressed in

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

Section 2, a Hub-and-Spoke architecture is particularly suited for that as it is scalable and reduces maintenance costs.

- **Autonomy and efficiency**: the system must be able to run autonomously by updating data on regular basis. Big Data technologies [21] can be considered to ensure it is able to process and store data efficiently.
- **Robustness**: the system should be designed such that a failure in any of its components does not propagate to the others. For instance, a failure of the ETL facilities due to an Input/Output exception should stop the data propagation in order to avoid services receiving corrupted data.
- **Reliability**: the system and its services should be tested for an adequate period before making it available to the final users. Tests should stress the capacity limits of the entire infrastructure, e.g. in terms of CPU, memory and disk space. Interfaces should be tested by means of usability studies.

After an initial investigation, we found that there are no comprehensive commercial solutions able to support all the stages of the proposed methodology. At the end, the choice about the specific technologies to employ was made by considering both the features of the various alternative partial solutions found and prior knowledge of the people of the project team. We employed a combination of Java and Coffee Scripts to develop the ETL facilities, the "SWEB" technology to instantiate the Hub, and a combination of ElasticSearch, Angular and C# to develop the services running on the spokes. In particular, the "SWEB" technology was developed by the Knowdive research group. It offers APIs and a Graphical User Interface (GUI) supporting all the necessary tasks including data modelling, vocabulary management, data integration and querying. It internally represents data as knowledge graphs. See for instance the work described in [22, 23] as partial examples of research applications of this work. The usage of "SWEB" also proves the ability of the University of Trento to apply and value its own research outcomes.

Legal and security challenges

Legal and security challenges include the difficulty to comply with Intellectual Property Rights (IPR), licensing and privacy concerns and guarantee secure access to data.

In terms of IT security, we selected technologies by making sure that they satisfy security levels demanded by Italian law. Our IT staff constantly ensures that adequate security measures are in place. System components

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

are secured and not accessible from outside of the University network. Access to them is granted to administrators only. Regular backups guarantee for the data integrity.

To protect the privacy of users, the system must be compliant with the General Data Protection Regulation (GDPR). In this respect, our Legal staff gives us constant advice. In designing and developing the system and the services, we followed privacy-by-design principles. We adopted the design strategies proposed by Hoepman [24]. For instance, the SEPARATE strategy recommends separating data in order to prevent unwanted correlations. By indexing data in different ElasticSearch indexes in different spokes, we make sure each spoke receives only the data that is strictly relevant for the digital service it supports. Such strategies have been recently suggested by the European Data Protection Supervisor (EDPS) as good example of approach that can be followed for identifying measures to implement privacy requirements[xiii].

In terms of IPR and licensing, we promote and support the download of Open Access publications with Creative Commons licenses through the institutional portal we developed. Users found particularly useful to access the full text directly from their personal page on the new institutional portal instead of browsing the institutional repository[xiv].

User-related challenges

We believe that one of major risks to be managed is failing to meet user expectations, both in terms of functionalities offered and time of delivery. A possible way to mitigate this risk is ensuring proper and constant communication with them. Users should be involved in all stages of the work and periodically informed about the progress. In particular, they should be made well aware that data offered by the services is only the visible tip of a huge invisible iceberg represented by the data stored in the original data sources. Therefore, imperfections in data exposed by the services is not necessarily due to a fault of the data integration IT infrastructure, but it may be rather due to incorrect data at the source. Such imperfections should be fixed in the original sources such that, by propagation, all systems benefit from the improved data quality. To ensure proposed solutions are well received by the entire community, we set up an institutional board that includes representatives from the various academic departments, as well as of the students, the relevant managerial staff and rector's delegates. Important decisions are taken during board meetings.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

5. Summary

The digital transformation poses new challenges for universities. In particular, the capability to offer complete, up-to-date and consistent information to their users across different communication channels and digital services is essential for universities. We explained how we acquired this capability in Trento by designing and implementing an IT infrastructure based on the Hub-and-Spoke paradigm that explicitly addresses data fragmentation and diversity, and by following a methodology that incorporates typical LIS practices such as authority and vocabulary control. We described the vision and status of the Digital University initiative running in Trento and briefly illustrated what we learned so far. We will continue our work by extending the spectrum of data and services offered and collaborate with other universities that would like to challenge themselves to reach similar goals.

References

1. McDonald, M. P., and Rowsell-Jones, A. (2012). The Digital Edge, Exploiting Information and Technology for Business Advantage. Gartner, Inc.
2. Gartner (2014). Gartner Says One Third of Fortune 100 Organizations Will Face an Information Crisis by 2017. Available at: http://www.gartner.com/newsroom/id/2672515
3. Maltese, V, and Giunchiglia, F (2017). Foundations of Digital Universities. Cataloging & Classification Quarterly, 55(1), 26-50.
4. Buchanan, L, and O'Connell, A (2006). A brief history of decision making. Harvard Business Review, 84(1), 32-40.
5. Brynjolfsson, E, Hitt, L M, and Kim, H H (2011). Strength in numbers: How does data-driven decision making affect firm performance? Working Paper, Sloan School of Management, MIT, Cambridge, MA
6. Lenzerini, M (2002). Data integration: a theoretical perspective. In: twenty-first ACM SIGMOD-SIGACTSIGART symposium on principles of database systems. ACM, 233-246.
7. Watson, H J, and Wixom, B H (2007). The current state of business intelligence. Computer, 40(9), 96-99.
8. Denning, P J (2003). Computer science. Chichester (UK): John Wiley and Sons Ltd.
9. Buckland, M (1996). Documentation, information science, and library science in the USA. Information processing & management, 32(1), 63-76.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

10. O'Neill, E T (2011). FRBR: Functional requirements for bibliographic records. Library resources & technical services, 46(4), 150-159.
11. Zeng, M L, Žumer, M, and Salaba, A. (2011). Functional requirements for subject authority data (FRSAD): a conceptual model (Vol. 43). IFLA series on bibliographic control. Walter de Gruyter.
12. International Organization for Standardization (2011). ISO 2596-1:2011. Information and documentation-Thesauri and interoperability with other vocabularies: Part 1: Thesauri for information retrieval. ISO.
13. Sompel, H V D, Nelson, M L, Lagoze, C, and Warner, S (2004). Resource harvesting within the OAI-PMH framework. D-Lib Magazine, 10 (12).
14. Börner, K, Conlon, M, Corson-Rikert, J, and 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.
15. Tran, E, and Scholtes, G (2015). Open Data Literature Review. University of California, Berkeley School of Law, 17.
16. Waller, M A, and Fawcett, S E (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77-84.
17. Hopkins, B, Owens, L, Goetz, M, Gualtieri, M, and Keenan, J (2015). Deliver On Big Data Potential With A Hub-And-Spoke Architecture. Forrester Research.
18. Gómez-Pérez, A (2001): Evaluation of ontologies. International Journal of intelligent systems. 16(3), 391-409.
19. Chatterjee, U, Giunchiglia, F, Madalli, D P, and Maltese, V (2016). Modeling Recipes for Online Search. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pp. 625-642, Springer International Publishing.
20. Maltese, V, and Giunchiglia, F (2016). Search and Analytics Challenges in Digital Libraries and Archives. Journal of Data and Information Quality, 7(3), 10-12.
21. Chen, M, Mao, S, and Liu, Y (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209.
22. 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.

This is a preprint of an article whose final and definitive form has been published in the **Cataloging & Classification Volume 56, 2018 - Issue 7** - https://doi.org/10.1080/01639374.2018.1504847

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

24. Hoepman, J (2014). Privacy design strategies. ICT systems security and privacy protection, Springer Berlin Heidelberg, 446-459.

[i] The EUPRIO 2018 Conference website: https://www.euprio.eu/conference-seville/sevilla-conference-2018/
[ii] A Wikipedia page describing the HUB&Spoke paradigm: https://en.wikipedia.org/wiki/Spoke%E2%80%93hub_distribution_paradigm
[iii] The Linked Universities initiative website: http://linkeduniversities.org/
[iv] The VIVO project full documentation: https://wiki.duraspace.org/display/VIVO
[v] The W3C standard website for RDF: https://www.w3.org/RDF/
[vi] The W3C standard website for SPARQL: https://www.w3.org/TR/rdf-sparql-query/
[vii] The W3C standard website for OWL, section describing sameAs: https://www.w3.org/TR/owl-ref/#sameAs-def
[viii] The Digital University initiative website: http://www.unitn.it/en/DU/info
[ix] The Knowdive research group website: http://disi.unitn.it/~knowdive/
[x] The JSON standard website: https://www.json.org/
[xi] A Wikipedia page describing REST: https://en.wikipedia.org/wiki/Representational_state_transfer
[xii] The Elasticsearch product website: https://www.elastic.co/products/elasticsearch
[xiii] The EDPS Preliminary Opinion on Privacy by Design webpage (31 May 2018): https://edps.europa.eu/data-protection/our-work/publications/opinions/privacy-design_en
[xiv] The institutional repository of publications of the University of Trento: https://iris.unitn.it/