Journal of Telecommunications and the Digital Economy

Addressing Digital Transformation in Universities

How to Effectively Govern, Trust and Value Institutional Data

Vincenzo Maltese
Dissemination and Evaluation of Research Results Division,
University of Trento, Italy

Abstract: In facing digital transformation challenges, universities need to set up their data governance strategies. They include effective solutions to trace and value data about key assets (such as researchers, publications, courses, research projects) scattered across multiple legacy IT systems. As part of an overall solution to deal with the unavoidable data fragmentation and diversity, we provide the complete code of a simple and very efficient framework that can be employed by universities to develop their own knowledge graph, offering a comprehensive picture of the strategic data of the university, such that it can be consistently exploited by different digital services.

Keywords: Digital Transformation, Data Integration, Knowledge graphs, Vocabularies.

Introduction

In pursuing their missions along the three pillars of education, research and societal impact, universities need to find their own way to address the challenges increasingly posed by digital transformation. The term digital transformation usually indicates a set of technological, cultural, organizational, social, creative and managerial changes (McDonald *et al.*, 2012). Digital transformation goes beyond the simple adoption of new technologies and makes it possible to provide services, supply goods, exhibit live experiences, and find, process and make accessible large amounts of content regardless of the real availability of resources, pervasively creating new connections between people, places and things.

Digital transformation in higher education institutions is about the development of new, more advanced and effective methods and practices in pursuit of higher education's mission (Alenezi, 2021). Even though it brings new opportunities, digital transformation also poses new challenges for Communication and IT departments of universities (Maltese, 2018a). Recent studies (Safiullin & Akhmetshin, 2019; Gafurov *et al.*, 2020; Marks & Al-Ali, 2022)

http://doi.org/10.18080/jtde.v12n1.808

confirm that universities are not yet prepared, in terms of vision, competency, infrastructures, data strategies and digitalization of their services.

Our work focuses on *digital information challenges*. Universities need to provide to their stakeholders detailed information about a variety of key assets, such as professors, researchers, employees, publications, courses, and research projects. It is, however, difficult for universities to present a complete, up-to-date and coherent picture about them across the different digital communication channels and services employed. For example, it may happen that a certain person is an associate professor according to the human resources system (the main authority for such data), a research fellow on the main institutional portal (the portal is outdated), and a post-doctoral researcher on the department website (the website is not only outdated, but it uses different terminology with respect to the institutional portal).

The root of this difficulty lies in the inherent complexity of the IT university ecosystem (Maltese, 2018b) and it is common to many other large-scale organizations (Gartner, 2014). The diversity of IT systems is actually needed to target specific business processes and key assets with confined responsibility. As a consequence, data fragmentation and diversity (that progressively increase with the number of IT systems employed and the growth of data) bring about a sort of entropic effect where: 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.

Solutions to this problem can be altogether referred to as *data governance strategies*. We report our experience matured during research (Giunchiglia *et al.*, 2012b; Giunchiglia *et al.*, 2014; Maltese & Giunchiglia, 2016; Maltese & Giunchiglia, 2017) and innovation (Maltese, 2018b; Giunchiglia *et al.*, 2022) projects conducted in universities and provided further insights that have been presented during a series of invited talks (Maltese, 2017; 2018a; 2023a; 2023b; 2023c).

Maltese & Giunchiglia (2017) proposed a general solution to address this problem in universities. It stands in addressing *data diversity* via the adoption of well-established Library & Information Science methodologies and tools to curate data and metadata quality, and in addressing *data fragmentation* via the adoption of data integration methodologies and tools.

Maltese (2018b) provides the description of the system architecture, the tools and the digital services that were developed at the University of Trento in Italy in the context of the Digital University initiative and that constitute the first implementation of the general solution. The infrastructure follows the Hub-and-Spoke paradigm. The Hub is an IT system that collects data extracted from various data sources and encodes it as a knowledge graph. This is achieved by means of Extract, Transform and Load (ETL) facilities. In the Extract phase, data is selected from relevant legacy IT systems. In the Transform phase, data diversity is addressed by codifying data uniformly. In the Load phase, data fragmentation is addressed by collecting and pulling together into the Hub data about the same entity (e.g., a single person or a single publication). The knowledge graph provides centralized access to a number of Spokes, each of them being a new IT system expressly developed to support a different digital service. We described the challenges that typically arise (Maltese & Giunchiglia, 2016) and how we addressed them in Italy and in Mongolia (Giunchiglia *et al.*, 2022). Similar issues have been discussed by Rodríguez & Bribiesca (2021), Tungpantong *et al.* (2021), Esmailzadeh *et al.* (2022), Gkrimpizi & Peristeras (2022) and Sułkowski (2023).

The main contribution of this paper is the description and the complete source code of a new data integration framework that we developed in 2021, and that is now publicly available on GitHub (https://github.com/vinmal74/DU). It entirely substitutes the one employed in the first version of the Digital University system developed between 2017 and 2018. It supports engineers in the creation of a multilingual knowledge graph from data extracted from multiple sources. Entirely developed in Java (the previous one required several different technologies, including Java, Scala and Coffee scripts), it makes the development of the ETL facilities much simpler. By changing the entity matching algorithm (that is necessary to detect and merge duplicates) and the data structures employed, it allowed us to overcome the technical challenges described in Giunchiglia *et al.* (2022), thus reducing the time needed to create the knowledge graph by three orders of magnitude with respect to the previous version. In terms of computational complexity, the new algorithm is linear in the number of entities to be integrated, while the previous one was quadratic in the number of entities. It is faster also because of the data structures employed (hash maps), stored entirely in RAM memory (the previous version operated entirely on databases stored in the file system).

In the rest of the paper, we summarize the state of the art, and recall the system architecture and methodology employed, as illustrated in our previous work. We continue with the main contribution of this paper, that is the source code of the new framework for the creation of the knowledge graph at the core of the Digital University solution. Our aim is to provide the methodology and tools such that other universities can replicate our work. Therefore, we provide a demonstrative example of data sources and the ETL code necessary to create the corresponding knowledge graph. The source code, the example and the ETL code are fully available on GitHub. We also illustrate how the knowledge graph can be consistently used by multiple digital services. Finally, we summarize the work done and the future work.

State of the Art and Related Work

Several research communities traditionally address data fragmentation and diversity (Maltese *et al.*, 2009). In the following, we focus on the solutions proposed by Business Intelligence (BI) and Library & Information Science (LIS).

The primary purpose of BI is to support decision-making in organizations (Buchanan & O'Connell, 2006). Data-driven decision-making refers to the practice of basing decisions on the analysis of data rather than purely on intuition (Brynjolfsson *et al.*, 2011). 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 the data (Lenzerini, 2002). 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 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 (Watson & Wixom, 2007).

Library Science is traditionally concerned with archiving texts and organizing storage and retrieval systems to give efficient access to texts (Denning, 2003). LIS is the technical and technological innovation of Library Science that employs information technology for documentation and library services (Buckland, 1996). 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 (O'Neill, 2011). 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 (O'Neill, 2011). Unique headers include names and alphanumeric identifiers. Similarly, vocabulary control enforces the usage of standard terms to unambiguously refer to each subject (Zeng *et al.*, 2011). In controlled vocabularies, standard terms are arranged hierarchically from broader to narrower terms (ISO 2596-1:2011). Altogether, the adoption of these practices enables 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, such as the OAI-PMH framework (Sompel *et al.*, 2004) and by adopting solutions to map equivalent concepts in different knowledge organization systems (ISO 2596-1:2011; Giunchiglia *et al.*, 2009; Maltese *et al.*, 2010; Giunchiglia *et al.*, 2012a).

A few initiatives have provided solutions to support storing, searching, browsing, visualizing and sharing scholarly data. VIVO (Börner *et al.*, 2012) relies on Semantic Web technologies to represent and store data in the RDF standard model (https://www.w3.org/RDF/) and retrieve it using the SPARQL query language (https://www.w3.org/TR/rdf-sparql-query/). However, it has been observed that these initiatives offer limited support to tackle data diversity and data fragmentation (Maltese & Giunchiglia, 2017). In fact, they do not provide effective entity matching tools and methodologies to effectively control and enforce terminology.

Our approach is compliant with other solutions designed for universities, such as VIVO, and for digital libraries, such as DSPACE (Smith *et al.*, 2003). For instance, a converter can be easily developed to translate our knowledge graph into the VIVO model and ontology, so that it can be exploited by VIVO applications, such as the VIVO portal. Our framework makes the creation of the knowledge graph simple and very efficient.

The System Architecture

The system architecture adopted in Trento (Figure 1) was first introduced in Maltese (2018b) and described further in Giunchiglia *et al.* (2022). The knowledge graph is built by reusing data that becomes available through ETL facilities, and it is employed in a Hub-and-Spoke architecture. Each spoke supports a different digital service. The idea is that new Spokes are added incrementally whenever there is a need for a new service which cannot be provided by the existing Spokes.

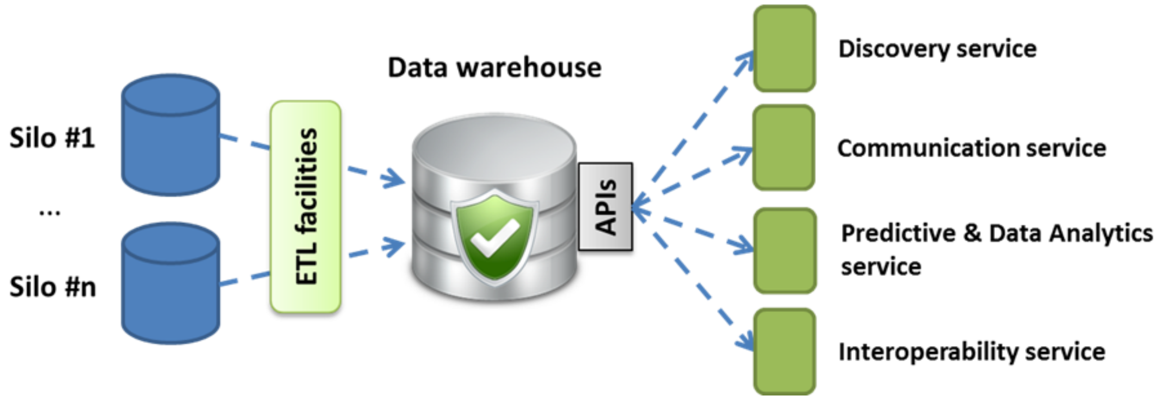


Figure 1. The system infrastructure of Digital Universities

This architecture was chosen as it represents a more efficient and scalable alternative to point-to-point communication in that the number of connectors between IT systems is reduced drastically, thus reducing complexity and maintenance costs (Hopkins *et al.*, 2015).

The Hub collects data extracted from various data sources (Extract), encodes data according to a uniform model and terminology (Translate), and creates a knowledge graph through an integration framework (Load). Through dedicated Application Programming Interfaces (APIs), the Spokes get access to the knowledge graph stored in the Hub.

Overall, the Hub fulfils the following requirements (Maltese, 2018b).

- **The Hub provides centralized access to data** natively stored in the heterogeneous data sources (different schema, model and format) managed by legacy IT systems. This separation of duties is necessary to ensure that legacy systems can continue to function as usual, thus benefitting from advantages (contained costs, dedicated business processes, focused data, dedicated users and confined responsibilities) that come from their vertical end-user applications. Relevant data about the key entities that are necessary to support the centralized services is reused in the Hub by means of ETL facilities. They ensure that data about the same entity extracted from multiple sources is appropriately collected, transformed, merged and correlated. Especially, entity matching (e.g., Wang *et al.*, 2011) and merge facilities are essential to avoid the presence of duplicates.
- **The Hub supports knowledge and language localization**, in that the knowledge graph is built according to a local customized data model and terminology. Localization can take place starting from a reference data model (the knowledge) and vocabulary (the language) designed specifically for universities (Maltese, 2018b). Their main purpose is to provide a common core of entity types, properties and terminology in multiple languages necessary to fulfil typical services of a university and to favour interoperability among them, similarly to what is done by VIVO. Simultaneously, the different needs across the globe demand the capability of the system to support their customization and extension as required locally by the digital services of a certain university.
- **The Hub supports the development of centralized services** via dedicated APIs that provide access to the knowledge graph. APIs support the development of university services on the Spokes such that they can consistently query the Hub and exploit the same content, i.e., the knowledge graph. They include: (a) *Discovery services* supporting browsing and search (Giunchiglia *et al.*, 2014); (b) *Communication services* conveying information to stakeholders uniformly and consistently across different communication channels (Maltese, 2018b); (c) *Predictive & data analytics services* supporting decision-making processes (Waller & Fawcett, 2013; Brdesee, 2021); (d) *Interoperability services* supporting the import/export of data from/to existing standards, such as the publication of Open Data (Tran & Scholtes, 2015), or according to the VIVO model and ontology, or to answer queries across federated universities.

Among other things, in our previous work (Maltese, 2018b) we described how we 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. Data sources and system components are secured and not accessible from outside of the University intranet. Access to them is granted to administrators only. Data is accessed exclusively via database views expressly arranged to provide access to relevant data only. Among other things, this makes system maintenance easier in that such views can be seen as *contracts* that cannot be violated even in the case that the data source changes, e.g., because of an update of the corresponding IT system. Regular backups guarantee data integrity.

To protect the privacy of users, and to be compliant with the General Data Protection Regulation (GDPR), in designing and developing the system and the services we followed well-established privacy-by-design principles (Hoepman, 2014), suggested also by the European Data Protection Supervisor (2018). Our privacy policies are publicly available. Only relevant and non-sensitive data is managed. Data is stored in separate indexes in order to prevent unwanted correlations. Each Spoke receives only the data that is strictly relevant for the digital service it supports. In terms of IPR, we promote and support Open Science principles by allowing the download of scientific publications of our researchers with Creative Commons licenses through the institutional portal we developed.

The Methodology

The methodology, introduced in Maltese & Giunchiglia (2017) and refined in Giunchiglia *et al.* (2022), defines an iterative process composed of sequential steps, briefly outlined below, which are followed every time a new digital service needs to be designed and developed. In our previous work, we illustrated its advantages that include scalability, cost-effectiveness, and facilitated compliance with legal constraints.

Step 1. Collecting service requirements. It consists of collecting the requirements of the new service in terms of functionalities, target users and necessary data.

Step 2. Knowledge localization. The reference data model, providing the schema which is enforced to store the knowledge graph in the Hub, is adapted to local needs. It is constituted by entity types and properties necessary to describe typical key entities of universities, such as people, courses, publications, dissertations and research projects. Chatterjee *et al.* (2016) presents a methodology that can be followed to design the data model in a given domain. It should include identifiers, i.e., those properties necessary to identify unequivocally an entity of a certain type such that entity matchers can work properly (Bouquet *et al.*, 2007). Knowledge adaptation means adding or specializing entity types and properties that are necessary to support the new service.

Step 3. Language localization. The controlled vocabulary is adapted to local needs. We employ well-established LIS methodologies for vocabulary development (Maltese, 2018b). For instance, the vocabulary should provide the terminology necessary to describe the various positions occupied by people (e.g., full professor, associate professor, researcher), the kinds of publications (e.g., journal article, conference paper), the status of a research project (e.g., submitted, approved). Language adaptation means adding or specializing concepts, selecting preferred terms from the vocabularies or adding new languages that are necessary to support the new service. This includes handling lexical gaps (Giunchiglia *et al.*, 2018), i.e., concepts which do not have a precise translation in the target language. We address language diversity by representing knowledge as language-independent concepts whose meaning is approximated in each language by means of terms that are the closest in meaning.

Step 4. Data hunting. The legacy IT systems are assessed in order to identify the possible sources for the data required by the service. The following cases can arise: (a) there is only one system that can provide them; (b) multiple systems, possibly maintained by different departments, can provide part of them, which can eventually partially overlap or even be in conflict; or (c) existing systems cannot provide all of them. In the latter case, it is necessary to develop new IT systems able to complete missing data.

Step 5. Building the knowledge graph. ETL facilities are implemented in order to Extract and Translate data according to the localized knowledge and language, and to Load them into the Hub. Mechanisms to resolve conflicts in data may include authority (based on the ordering of importance of the sources) or voting (based on the majority of the sources) schemes (Dong & Naumann, 2009). Overlaps are handled through entity matching and merging techniques. This task requires an adequate infrastructure able to semi-automate the process and to keep the Hub aligned with the sources, by running ETL facilities regularly (e.g., daily). Especially, human intervention is required to fix mistakes in data (whenever possible, they should be fixed in the data sources), accommodate for missing terms in the controlled vocabulary (thus requiring an extension of the vocabulary) and when the schema of the data sources changes (e.g., an attribute was supposed to have *n* possible values and the (n+1)th value appears). Fixes are recorded and applied automatically in the next updates (Giunchiglia *et al.*, 2021).

Step 6. Implementing the service. The service is implemented and deployed by accessing the knowledge graph data from the Hub via dedicated APIs.

The Digital University Framework

The framework we developed to support the creation of the multilingual knowledge graph is fully available at https://github.com/vinmal74/DU/tree/main/src/Hub.

The framework is simple in that it is entirely developed in Java, it is constituted by less than 300 lines of code, and it is based on the well-known object-oriented programming paradigm. The Entity Relationship model (Chen, 1976) is employed to represent the various entities and how they are interconnected. The ETL paradigm, which is typical of data warehousing approaches to data integration (El-Sappagh *et al.*, 2011), is employed to extract data from the original data sources, to convert them into entities, and to incrementally construct the knowledge graph. Such simplicity allows any programmer, with no specific knowledge of representation languages and Semantic Web technologies, to adopt it very quickly and easily.

The framework is very efficient for two reasons. The first is that the data integration algorithm is linear in computational complexity. In fact, it employs hash maps to store and retrieve the entities: insertion and retrieval in hash maps takes constant time. The second is that all data structures are stored in RAM memory to guarantee the maximum performance at runtime.

For instance, the knowledge graph of the University of Trento is currently constituted by around 225,000 entities, appropriately selected. Entity types are Person, Organization, Role, Course, Project, Thesis, Publications, and Files. The creation of the knowledge graph takes 2-3 minutes (depending on the network load, given that datasets are located on different servers) on a laptop equipped with an Intel Core i5-7200 dual core 2.50 GHz and 2.71 GHz, and 8 GB of RAM memory. The total memory usage is around 370 MB.

The framework consists of 10 Java classes. Figure 2 provides an exemplification of the data structures used to represent the knowledge graph. It shows three entitybases and three entities interconnected between them. The mapping between the classes of the framework to the standard W3C RDF schema (https://www.w3.org/TR/rdf-schema/) is trivial.

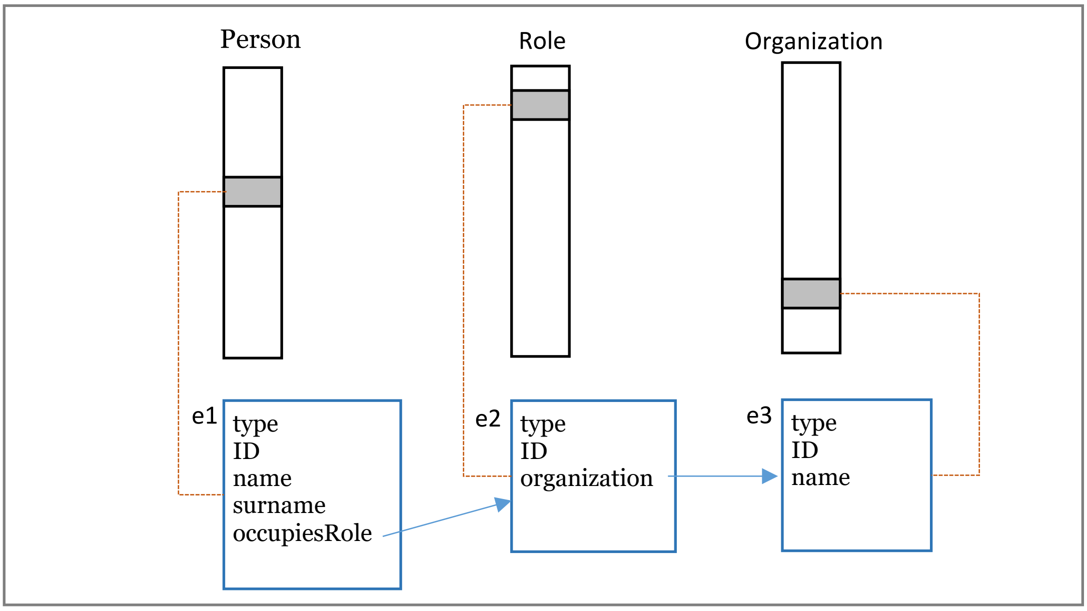


Figure 2. An exemplification of an entityStore

The knowledge graph is represented as a set of **entitybases** (Entitybase.java), one for each entity type required. Within each entitybase, we employ a HashMap. We represent entity types as an integer. We suggest that types could be encoded as constant values, e.g., Person = 0, Organization = 1, Role = 2, Course = 3. Each **entity** (Entity.java) is characterized by its type, a unique identifier (that we represent as a String) and a set of attributes.

Attributes (Attribute.java) are <name, value> pairs. The current framework supports three different types of attributes. Java classes can be extended to support additional types. **String attributes** (StringAttribute.java) are language independent attributes whose value is stored as a String; numbers and dates are converted into strings. **Relational attributes** (RelationalAttribute.java) represent relations between entities; in fact, their values (EntityValue.java) are <type, id> pairs, where type is the entity type and id is the identifier of the target entity. **Concept attributes** (ConceptAttribute.java) are language-dependent attributes whose values (ConceptValue.java) are stored as a Concept (that are simply represented as integers) used to codify values whose labels need to be read according to the languages used, e.g., in English and Italian.

In order to represent concepts in specific languages, it is necessary to define corresponding **vocabularies** (Vocabulary.java). Each vocabulary is characterized by a reference language and a list of **concepts** (Concept.java). Each concept is a triple <id, label, definition>. For example, the concept of researcher in English is given by the triple <56569, “researcher”, “(role) a person who conducts research activities”>, while in Italian it is given by <56569, “ricercatore”, “(ruolo) una persona che svolge attività di ricerca”>. Following the ISO 25964 standard for vocabulary representation (https://www.iso.org/standard/53657.html), the identifier must obviously be the same in all vocabularies. The definition is needed to keep track of the meaning of the labels.

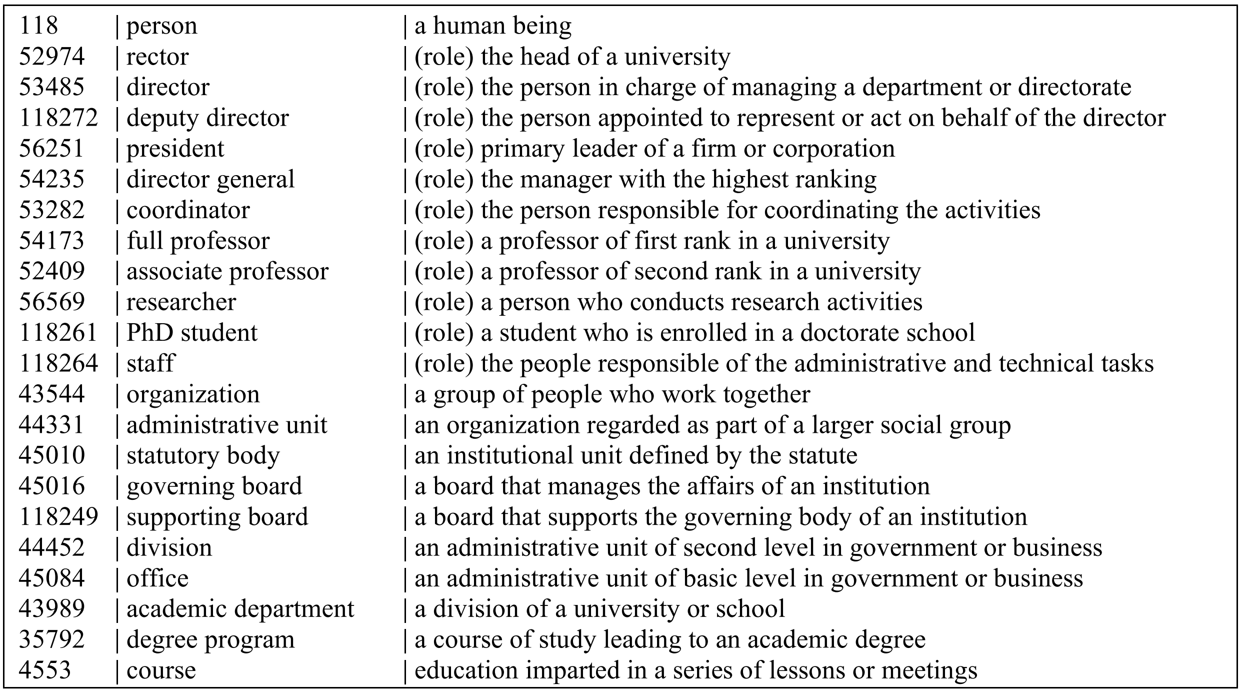
The current framework has two main limitations. The first is that it may need significant amount of RAM in case of datasets of huge size. For the purposes we envisioned in Trento (see Using the Knowledge Graph in Multiple Digital Services below), the RAM memory used is actually approximately 370 MB only. Such cheap usage of memory is possible because we only select relevant data to be extracted from the datasets. The second limitation is that it may require an extension of the entity-matching libraries in case not all sources already provide unique identifiers for all entities or in case similar entities are stored in different datasets with different identifiers. We overcome this limitation by making sure that identifiers are always available for all entities, either as a single attribute, or as a result of a combination of multiple attributes.

The Demonstrative Example

Suppose we want to develop a university portal in two languages, English and Italian, whose functionalities have been identified by collecting requirements from various stakeholders. The local data model will have to define the various entity types and their attributes necessary to accommodate such requirements. For instance, it may establish that a Person must have name, surname, gender, email, phone and set of positions occupied in administrative units. For sake of simplicity, we assume that data sources have been already pre-processed (for instance, as a result of a job that runs daily in order to get up-to-date data) and available as CSV files (see https://github.com/vinmal74/DU/tree/main/src/data):

- **people.csv** contains the people affiliated to the University;
- **units.csv** contains the administrative units of the University;
- **types_of_units.csv** contains information about the types of administrative units;
- **positions.csv** contains information about the affiliations of each person;
- **types_of_positions.csv** contains information about the types of positions that can be appointed to people in the administrative units;
- **courses.csv** contains information about the courses offered.

The two vocabularies are stored in TXT files. Each row contains the identifier of the concept, the label and the definition in the corresponding language. They can be extended as needed. Figure 3 provides a fragment of the content in the English vocabulary.



```
118     | person               | a human being
52974   | rector               | (role) the head of a university
53485   | director             | (role) the person in charge of managing a department or directorate
118272  | deputy director      | (role) the person appointed to represent or act on behalf of the director
56251   | president            | (role) primary leader of a firm or corporation
54235   | director general     | (role) the manager with the highest ranking
53282   | coordinator          | (role) the person responsible for coordinating the activities
54173   | full professor       | (role) a professor of first rank in a university
52409   | associate professor  | (role) a professor of second rank in a university
56569   | researcher           | (role) a person who conducts research activities
118261  | PhD student          | (role) a student who is enrolled in a doctorate school
118264  | staff                | (role) the people responsible of the administrative and technical tasks
43544   | organization         | a group of people who work together
44331   | administrative unit  | an organization regarded as part of a larger social group
45010   | statutory body       | an institutional unit defined by the statute
45016   | governing board      | a board that manages the affairs of an institution
118249  | supporting board     | a board that supports the governing body of an institution
44452   | division             | an administrative unit of second level in government or business
45084   | office               | an administrative unit of basic level in government or business
43989   | academic department  | a division of a university or school
35792   | degree program       | a course of study leading to an academic degree
4553    | course               | education imparted in a series of lessons or meetings
```

Figure 3. Example of content of the vocabularies (in English)

Developing the ETL Facilities

In this section, we present a demonstrative toy example of how the knowledge graph can be built by implementing ETL facilities and by employing the framework. It is fully available on GitHub at https://github.com/vinmal74/DU/tree/main/src/ETL.

The main functionality offered by an entitybase is data integration, supported by the load method (see Entitybase.java). As from the example in Figure 4, suppose we extracted enough data to generate the entity e1. In loading the entity e1 in the entitybase E, if E already contains an entity e2 with the same identifier, the set of attributes of e1 are merged with those of e2, thus obtaining the entity e3; otherwise e1 is loaded in E as it is.

e1		**e2**		**e3**	
type	**= 118**	**type**	**= 118**	**type**	**= 118**
ID	**= 1000099**	**ID**	**= 1000099**	**ID**	**= 1000099**
Surname	= Sordi	Email	= alberto.sordi@unitn.it	Surname	= Sordi
Name	= Alberto	Phone	= 2005	Name	= Alberto
Phone	= 1000			Email	= alberto.sordi@unitn.it
				Phone	= [1000, 2005]

Figure 4. Example of data integration

Two attributes are considered to be different when the name or the value do not match. In the current implementation, both the entity matching and the attribute functions simply rely on the standard equality operator, but, according to the specific scenario, it could be a more complex similarity function (Köpcke & Rahm, 2010), e.g., to accommodate approximation of values.

Thus, a data integration pipeline can be designed as a set of ETL facilities where for each data source a dedicated facility extracts data (E), translates it into a set of entities (T), and loads each of them in the corresponding entitybase (L). Given that the load function is characterized by O(1) computational complexity, the complexity of the ETL algorithm is O(n), where n is the number of entities identified in the data sources.

In the following, we describe the code of the data integration pipeline. Individual entitybases are stored in an **EntityStore** (EntityStore.java). We implemented them as an array. The toy example requires four entitybases: EB[0] for Person, EB[1] for Organization, EB[2] for Role, EB[3] for Course.

The **data integration pipeline** (ETL.java) contains the main method. It creates the English and Italian vocabularies by loading the two TXT files that contain the <id, label, definition> triples, and initializes the EntityStore. Finally, it launches four different ETL facilities to process the CSV files with the data sources and to incrementally construct the knowledge graph. They can be executed in any order, thus always obtaining the same result.

The first ETL facility (People.java) processes people.csv. Below we exemplify how, for the first row of people.csv, it creates one entity of type person to be loaded in EB[0]. Here 118 and 90013 are the concept IDs for "person" and "male", respectively, in the vocabularies. Class is the attribute that can be used to specialize the type, that in this case remains "person".

type	**= 118**
ID	**= 1000099**
Class	= 118
Surname	= Sordi
Name	= Alberto
Gender	= 90013
Email	= alberto.sordi@unitn.it
Phone number	= 1000

The second ETL facility (Units.java) processes units.csv. Below we exemplify how, for the first two rows of units.csv, it creates two entities of type organization to be loaded in EB[1]. Here, 43544, 44834 and 45016 are the concept IDs for “organization”, “university” and “academic senate”, respectively, in the vocabularies. The latter two are taken from types_of_units.csv. The Class attribute specializes the type “organization”.

type	**= 43544**
ID	**= UNIT00001**
Class	= 44834
Name	= University of Trento

type	**= 43544**
ID	**= UNIT000002**
Class	= 45016
Name	= Academic Senate
Part of	= (1, UNIT00001)

The third ETL facility (Positions.java) processes positions.csv. Below we exemplify how, for the first row of positions.csv, it creates three entities. The first entity is of type person to be loaded in EB[0], and corresponds to the same person with ID 1000099 created above; it is therefore merged with the first version of the same entity previously loaded in EB[0], thus adding the relation Occupies Role. The second entity is of type organization to be loaded in EB[1], and corresponds to the same organization with ID 43544 created above; it is therefore merged with the first version of the entity loaded before in EB[1], but no additional attributes are added. The third entity is of type role to be loaded in EB[2], where 118247 is the concept ID for “role” and the type of position occupied OTHEXT001 is converted into 118264, which is the concept ID of “other staff” (see types_of_positions.csv) that becomes the value of Class. We represent roles similarly to Jureta *et al.* (2007).

type	**= 118**
ID	**= 1000099**
Occupies Role	= (2, UNIT000002_118264)

type	**= 43544**
ID	**= UNIT000002**

type	**= 118247**
ID	**= UNIT000002_118264**
Class	= 118264
Organization	= (1, UNIT000002)

The fourth ETL facility (Courses.java) processes courses.csv. Below we exemplify how, for the first row, it creates three entities. The first is of type person to be loaded in EB[0]. The second is of type organization to be loaded in EB[1]. The third is of type course to be loaded in EB[3], where 4553 is the Concept ID for “course”.

```
type           = 118
ID             = 1000313

type           = 43544
ID             = UNIT08624

type           = 4553
ID             = 90065
Class          = 4553
Name           = Administrative Law
Degree program = Law (LM5)
Department     = (1, UNIT08624)
Professor      = (0, 1000313)
```

Using the Knowledge Graph in Multiple Digital Services

Once the knowledge graph has been created, it needs to be stored somewhere. In Trento, we store it in Elasticsearch indexes (https://www.elastic.co/what-is/elasticsearch), a distributed, free and open search and analytics engine for all types of data that offers simple, very efficient and scalable REST APIs to store and query data. On top of Elasticsearch, we then developed an additional layer of RESTful APIs that are used by the various digital services.

To export the knowledge graph, the EntityStore offers the toJSON function that converts an entity into a JSON object. It takes as input the identifier of the entity, its type (to identify the entitybase in which it is contained), the vocabulary to be used to translate concept attributes (for instance, the Italian vocabulary), and the depth of the knowledge graph to be taken, i.e. the maximum number of relations to be followed. As an alternative, you can directly use the get functions offered by the Java classes.

So far, in Trento we have designed and developed four digital services that access to the same knowledge graph.

The **institutional portal** (https://webapps.unitn.it/du/en) is a public communication service that offers a comprehensive webpage (from the integration of 7 different data sources) in English and Italian for each of the University members, academic departments, governing bodies and administrative units. University members include academic staff (professors, researchers, PhD students), administrative and technical staff, and university executives. It is visited by around one million users per year.

The **institutional dashboard** is a data analytics service providing insights about the quality of research conducted by the faculty members with a focus on publications and research projects. It provides statistics and interactive graphs useful to examine trends, strengths, and points of improvement. Access is reserved to University members only via credentials.

Dedicated APIs have been developed for the **publication of Open Data** on the regional (https://dati.trentino.it/organization/universita-di-trento), Italian (https://www.dati.gov.it) and European (https://data.europa.eu) data portals. With this interoperability service, we comply with national guidelines about sharing public sector information.

The **University Mobile App** (https://unitrento.app/) is a communication service that has been developed by the IT staff of the University for its students. The knowledge graph is one of the data sources used and is accessed through dedicated APIs.

Conclusions

Digital transformation poses new challenges for universities. They need to tune their strategies for effective data governance and identify efficient solutions to trace and value information about their key assets scattered across multiple IT systems. As part of an overall solution to deal with the unavoidable data fragmentation and diversity, we illustrated the work done in Trento where we designed and implemented an infrastructure based on the Hub-and-Spoke paradigm. In this paper, we presented the new framework and the ETL facilities that we developed in 2021 to construct our knowledge graph more efficiently and easily than in the first version. The knowledge graph is used consistently by different digital services. We hope that the source code provided here can be of inspiration and can be employed by other universities to develop their own knowledge graphs and digital services.

References

Alenezi M. (2021). Deep Dive into Digital Transformation in Higher Education Institutions. *Education Sciences*, *11*(12), 770. https://doi.org/10.3390/educsci11120770

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. http://dx.doi.org/10.1007/978-3-031-79435-3

Bouquet, P., Stoermer, H., & Liu, X. (2007). Okkam4P: A Protégé Plugin for Supporting the Re-use of Globally Unique Identifiers for Individuals in OWL/RDF Knowledge Bases. In Proceedings of the Fourth Italian Semantic Web Workshop (SWAP), Bari, Italy, December 18-20, 2007, CEUR Workshop Proceedings, ISSN 1613-0073, online https://ceur-ws.org/Vol-314/41.pdf

Brdesee, H. (2021). A divergent view of the impact of digital transformation on academic organizational and spending efficiency: A review and analytical study on a university E-service. *Sustainability*, *13*(13), 7048. https://doi.org/10.3390/su13137048

Brynjolfsson, E., Hitt, L. M., & 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. https://doi.org/10.2139/ssrn.1819486

Buchanan, L., and O'Connell, A. (2006). A brief history of decision making. *Harvard Business Review*, *84*(1), 32–40.

Buckland, M. (1996). Documentation, information science, and library science in the USA. *Information processing & management*, *32*(1), 63–76.

Chatterjee, U., Giunchiglia, F., Madalli, D. P., & 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. http://dx.doi.org/10.1007/978-3-319-48472-3_37

Chen, P. P. S. (1976). The entity-relationship model—toward a unified view of data. *ACM transactions on database systems*, *1*(1), 9–36. http://dx.doi.org/10.1145/320434.320440

Denning, P. J. (2003). *Computer science*. Chichester (UK): John Wiley and Sons Ltd.

Dong, X. L., & Naumann, F. (2009). Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, *2*(2), 1654–1655. http://dx.doi.org/10.14778/1687553.1687620

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University-Computer and Information Sciences*, *23*(2), 91–104. https://doi.org/10.1016/j.jksuci.2011.05.005

Esmailzadeh, H., Mafimoradi, S., Hemmati, A. R., & Rajabi, F. (2022). Challenges and policy recommendations for IT governance in the University of Medical Sciences: a case study. *Journal of Health Administration*, *25*(3), 9–29. https://www.cabidigitallibrary.org/doi/pdf/10.5555/20230145914

European Data Protection Supervisor. (2018). Preliminary Opinion on privacy by design. European Union publication. Available from https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf

Gafurov, I. R., Safiullin, M. R., Akhmetshin, E. M., Gapsalamov, A. R., & Vasilev, V. L. (2020). Change of the Higher Education Paradigm in the Context of Digital Transformation: From Resource Management to Access Control. *International Journal of Higher Education*, *9*(3), 71–85. http://dx.doi.org/10.5430/ijhe.v9n3p71

Gartner. (2014). Gartner Says One Third of Fortune 100 Organizations Will Face an Information Crisis by 2017. http://www.gartner.com/newsroom/id/2672515

Giunchiglia, F., Soergel, D., Maltese, V., & Bertacco, A. (2009). Mapping large-scale knowledge organization systems. Proceedings of the 2nd International Conference on the Semantic Web and Digital Libraries (ICSD). http://hdl.handle.net/11572/75875

Giunchiglia, F., Maltese, V., & Autayeu, A. (2012a). Computing minimal mappings between lightweight ontologies. *International Journal on Digital Libraries*, *12*, 179–193. http://dx.doi.org/10.1007/s00799-012-0083-2

Giunchiglia, F., Maltese, V., & Dutta, B. (2012b). Domains and context: first steps towards managing diversity in knowledge. *Journal of Web Semantics*, *12–13*, 53–63. http://dx.doi.org/10.1016/j.websem.2011.11.007

Giunchiglia, F., Dutta, B., & Maltese, V. (2014). From Knowledge Organization to Knowledge Representation. *Knowledge Organization*, *41*(1), 44–56. http://dx.doi.org/10.5771/0943-7444-2014-1-44

Giunchiglia, F., Batsuren, K., & Freihat, A. A. (2018). One world–seven thousand languages. 19th international conference on computational linguistics and intelligent text processing. Lecture Notes in Computer Science, vol. 13396. Springer, Cham. https://doi.org/10.1007/978-3-031-23793-5_19

Giunchiglia, F., Bocca, S., Fumagalli, M., Bagchi, M., & Zamboni, A. (2021). iTelos--Purpose Driven Knowledge Graph Generation. arXiv preprint arXiv:2105.09418.

Giunchiglia, F., Maltese, V., Ganbold, A., & Zamboni, A. (2022). An Architecture and a Methodology Enabling Interoperability within and across Universities. In 2022 IEEE International Conference on Knowledge Graph (ICKG) (pp. 71–78). IEEE. http://dx.doi.org/10.1109/ICKG55886.2022.00017

Gkrimpizi, T., & Peristeras, V. (2022). Barriers to digital transformation in higher education institutions. In Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance (pp. 154–160). http://dx.doi.org/10.1145/3560107.3560135

Gómez-Pérez, A. (2001). Evaluation of ontologies. *International Journal of intelligent systems*, *16*(3), 391–409.

Hoepman, J. H. (2014). Privacy Design Strategies. In Cuppens-Boulahia, N., Cuppens, F., Jajodia, S., Abou El Kalam, A., & Sans, T. (eds), *ICT Systems Security and Privacy Protection*. SEC 2014. IFIP Advances in Information and Communication Technology, vol. 428. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-55415-5_38

Hopkins, B., Owens, L., Goetz, M., Gualtieri, M., & Keenan, J. (2015). Deliver On Big Data Potential With A Hub-And-Spoke Architecture. Forrester Research. Available from https://www.forrester.com/report/deliver-on-big-data-potential-with-a-hub-and-spoke-architecture/RES83303

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. https://www.iso.org/standard/53657.html

Jureta, I. J., Faulkner, S., & Kolp, M. (2007). An Agent-Oriented Enterprise Model for Early Requirements Engineering. *Handbook of Ontologies for Business Interaction*, *122*. http://dx.doi.org/10.4018/978-1-59904-660-0.ch008

Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, *69*(2), 197–210. http://dx.doi.org/10.1016/j.datak.2009.10.003

Lenzerini, M. (2002). Data integration: a theoretical perspective. In twenty-first ACM SIGMOD-SIGACTSIGART symposium on principles of database systems. ACM, 233–246. http://dx.doi.org/10.1145/543613.543644

Maltese, V. (2023a). "The Digital University", invited talk at the Data Scientia meeting, Trento, Italy. http://datascientia.disi.unitn.it/events/

Maltese, V. (2023b). "The data-driven university: how to effectively govern, trust and value university data and offer coherent digital services", invited talk at the Digitalization of Universities Conference. First ed.: https://university-conf.com/previous_university/; Second ed.: https://university-conf.com/previous_university_november2023/

Maltese, V. (2023c). "'Cataloguing' experts by competences: the Digital University project", invited talk at the Look beyond Subject indexing of non-book resources International Conference, Rome, Italy. https://www.aib.it/notizie/look-beyond/

Maltese, V. (2018a). "The data-driven university: how to effectively govern, trust and value university data to face 2020 challenges", invited talk at the European Association of Communication Professionals in Higher Education (EUPRIO) conference, Sevilla, Spain. https://www.euprio.eu/conference/xxx-euprio-conference

Maltese, V (2018b). Digital transformation challenges for universities: Ensuring information consistency across digital services. *Cataloging & Classification Quarterly*, *56*, 592–606. http://dx.doi.org/10.1080/01639374.2018.1504847

Maltese, V. (2017). "Digital University in Trento: Work Done and Next Steps", invited talk at the 4th Knowledge in Diversity Workshop, Trento, Italy. http://datascientia.disi.unitn.it/wp-content/uploads/2018/06/kid_workshop_2017_programme-2.pdf

Maltese, V., Giunchiglia, F., Denecke, K., Lewis, P., Wallner, C., Baldry, A., & Madalli, D. (2009). On the interdisciplinary foundations of diversity. At the first Living Web Workshop at the International Semantic Web Conference (ISWC). https://ceur-ws.org/Vol-515/livingweb2009_paper1.pdf

Maltese, V., Giunchiglia, F., & Autayeu, A. (2010). Save up to 99% of your time in mapping validation. Proceedings of On the Move to Meaningful Internet Systems, OTM 2010: Confederated International Conferences: CoopIS, IS, DOA and ODBASE, Part II (pp. 1044–1060). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-16949-6_28

Maltese, V., & Giunchiglia, F. (2016). Search and Analytics Challenges in Digital Libraries and Archives. *Journal of Data and Information Quality*, *7*(3), 10–12. http://dx.doi.org/10.1145/2939377

Maltese, V., & Giunchiglia, F. (2017). Foundations of Digital Universities. *Cataloging & Classification Quarterly*, *55*(1), 26–50. http://dx.doi.org/10.1080/01639374.2016.1245231

Marks, A., & Al-Ali, M. (2022). Digital transformation in higher education: a framework for maturity assessment. In COVID-19 Challenges to University Information Technology Governance (pp. 61–81). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-031-13351-0_3

McDonald, M. P., & Rowsell-Jones, A. (2012). *The Digital Edge, Exploiting Information and Technology for Business Advantage*. Gartner, Inc.

O'Neill, E. T. (2011). FRBR: Functional requirements for bibliographic records. *Library resources & technical services*, *46*(4), 150–159. http://dx.doi.org/10.5860/lrts.46n4.150

Rodríguez-Abitia, G., & Bribiesca-Correa, G. (2021). Assessing digital transformation in universities. *Future Internet*, *13*(2), 52. http://dx.doi.org/10.3390/fi13020052

Safiullin, M. R., & Akhmetshin, E. M. (2019). Digital transformation of a university as a factor of ensuring its competitiveness. *International Journal of Engineering and Advanced Technology*, *9*(1), 7387–7390. http://dx.doi.org/10.35940/ijeat.A3097.109119

Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley, R., & Walker, J. H. (2003). DSpace: An open source dynamic digital repository. *D-Lib Magazine*, *9*(1). http://dx.doi.org/10.1045/january2003-smith

Sompel, H. V. D., Nelson, M. L., Lagoze, C., & Warner, S. (2004). Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, *10*(12).

Sułkowski, Ł. (2023). *Managing the Digital University: Paradigms, Leadership, and Organization*. Routledge Studies in Organizational Change & Development Series Editor: Bernard Burnes. http://dx.doi.org/10.4324/9781003366409

Tran, E., & Scholtes, G. (2015). Open Data Literature Review. University of California, Berkeley School of Law. Available from https://www.law.berkeley.edu/wp-content/uploads/2015/04/Final_OpenDataLitReview_2015-04-14_1.1.pdf

Tungpantong, C., Nilsook, P., & Wannapiroon, P. (2021). A conceptual framework of factors for information systems success to digital transformation in higher education institutions. In 2021 9th International Conference on Information and Education Technology (ICIET) (pp. 57–62). IEEE. http://dx.doi.org/10.1109/ICIET51873.2021.9419596

Waller, M. A., & 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. http://dx.doi.org/10.1111/jbl.12010

Wang, J., Li, G., Yu, J. X., & Feng, J. (2011). Entity matching: How similar is similar. *Proceedings of the VLDB Endowment*, *4*(10), 622–633. http://dx.doi.org/10.14778/2021017.2021020

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer*, *40*(9), 96–99. http://dx.doi.org/10.1109/MC.2007.331

Zeng, M. L., Žumer, M., & Salaba, A. (2011). Functional requirements for subject authority data (FRSAD): a conceptual model (Vol. 43). IFLA series on bibliographic control. Walter de Gruyter. http://dx.doi.org/10.1515/9783110263787