An Architecture and a Methodology Enabling Interoperability within and across Universities

Fausto Giunchiglia
Department of Information Engineering and Computer Science (DISI)
University of Trento
Trento, Italy
fausto.giunchiglia@unitn.it

Vincenzo Maltese
Dissemination and Evaluation of Research Results Division
University of Trento
Trento, Italy
vincenzo.maltese@unitn.it

Amarsanaa Ganbold
Department of Information and Computer Science (DICS)
National University of Mongolia
Ulaanbaatar, Mongolia
amarsanaag@num.edu.mn

Alessio Zamboni
Department of Information Engineering and Computer Science (DISI)
University of Trento
Trento, Italy
alessio.zamboni@unitn.it

***Abstract*—We propose a general methodology and an infrastructure which allows to achieve interoperability within the same university and across universities. The former goal is achieved by *incrementally* defining and building a knowledge graph (KG) using data coming from multiple heterogeneous databases. Interoperability across universities is achieved by having a reference KG schema that each university can adapt to the local needs, but keeping track of the changes, and by natively supporting multilinguality. We achieve this latter requirement by exploiting a multilingual lexical resource containing more than one thousand languages and by seamlessly translating across the schemas and also (to some extent) across the data written in the local languages. The effectiveness of the proposed approach is proven by the services developed in the context of two different projects conducted in two universities in Italy and Mongolia.**

***Index Terms*—ontology, knowledge graph, semantic data integration**

I. Introduction

As institutions of higher education and research, universities need to offer a broad portfolio of services to their internal (e.g. students, professors, administrative staff) and external (e.g. other universities, companies) target users. Services are mainly centered on the production, custodianship, monitoring, fruition, and dissemination of knowledge. Thanks to the Public Sector Information (PSI) European Directive[1] and the Open Science principles[2], both administrative and research data have been recognized as a public asset. For instance, at the national level the Agency for Digital in Italy (AgID) promotes the adoption of practices and the development of infrastructures for the publication and reuse of public sector Open Data[3]. At European level, the Interoperable Europe programme[4] promotes the adoption of standard models, vocabularies, and formats to support cooperation among European public administrations.

Through their services, universities offer data about some fundamental key entities such as people, organizations, teaching programs and courses, research projects, papers, books, dissertations, and patents. One of the barriers to collecting and reusing these data is represented by their native fragmentation and heterogeneity [1], typically confined in separate information silos. In fact, universities employ a number of different IT systems to support their internal business processes such as library management, HR management, teaching support, research and technology transfer support, project management and fundraising, financial support, IT support, legal support, logistics, strategic planning, and so on. This difficulty is common to many other large-scale organizations. In fact, in 2014, Gartner said that a significant number of organizations, unable to organize themselves effectively, will experience an information crisis due to their inability to effectively value, govern and trust their enterprise information[5].

The solutions for data reuse adopted so far by universities are based on data integration and semantic interoperability approaches. We reviewed the institutional portals of the top 10

The work presented in this paper received funds by the "Digital NUM project (P2022-4222)" of the National University of Mongolia, and by the MIUR "Progetti di Ricerca di Rilevante Interesse Nazionale" (PRIN) 2017 – DD n. 1062.

[1]https://digital-strategy.ec.europa.eu/en/policies/psi-open-data

[2]See for instance the Open Science policy of the European Union https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science_en

[3]https://docs.italia.it/italia/daf/lg-patrimonio-pubblico/it/stabile/dati.html

[4]https://joinup.ec.europa.eu/collection/interoperable-europe/interoperable-europe

[5]https://www.gartner.com/en/newsroom/press-releases/2014-02-27-gartner-says-one-third-of-fortune-100-organizations-will-face-an-information-crisis-by-2017

Italian universities according to the CENSIS[6] Italian ranking and of the top 20 universities of the world according to THE – Times Higher Education [7] ranking. In Italy, we only found one remarkable solution at the University of Milan[8]. This solution has been designed by the Consortium of Italian universities and public institutions (CINECA) and is based on VIVO [2]. Several other universities worldwide use VIVO, e.g. the University of Florida [9]. Among the top universities worldwide, we found two notable cases, namely, the HKU Scholars Hub[10], based on a custom solution, and the John Hopkins portal[11], based on Elsevier Pure[12].

With respect to this work, the research described in this paper provides two main novel contributions.

An architecture and infrastructure which support interoperability both within and across universities via the construction of a knowledge graph (KG). Within a single university, the KG allows reusing data extracted from selected data sources to feed multiple services, like institutional portals and data analytics services. Across universities, it allows for the publication of Open Data in one or more standard formats, and the native integration of data across universities, thus enabling the construction of a worldwide university KG, a global KG integrating many local KGs. A key feature enabling the above is that the infrastructure natively supports the integration of multilingual data. An example of cross-universities service is global search for job opportunities.

A methodology to be followed iteratively to incrementally extend the infrastructure, the data model, the ontology and the KG for the incremental design and development of end user services, one service at a time. The methodology is scalable and cost-effective.

We provide two use cases - the first at the University of Trento in Italy[13], already in production, and the second under evaluation, at the National University of Mongolia in Mongolia[14] - that validate and prove the advantages of the proposed solution.

The paper is organized as follows. In Section II, we illustrate the architecture for semantic data integration. Section III describes the general methodology for service design and implementation. In Section IV, we describe the core ontology, i.e., the reference KG schema, and its design. Section V presents the two use cases, while Section VI describes some of the challenges we faced in the two projects. Finally, in Section VII we provide the conclusions.

[6]https://www.censis.it/
[7]https://www.timeshighereducation.com/
[8]https://expertise.unimi.it
[9]https://vivo.ufl.edu
[10]https://hub.hku.hk
[11]https://jhu.pure.elsevier.com/en
[12]https://www.elsevier.com/solutions/pure
[13]For the University of Trento, see https://webapps.unitn.it/du/en for a portal and https://dati.trentino.it/organization/universita-di-trento for the Open Data both implemented as services on top of the KG.
[14]For the National University of Mongolia, see http://du.num.edu.mn for the homepage of the services enabled by the solution describe here.

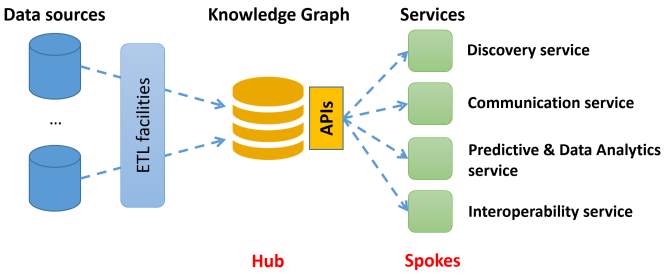

Fig. 1. The Hub-and-Spoke system architecture

II. The Architecture

The functional system architecture and infrastructure that we employ (Figure 1) was first introduced in [1] . The KG in Figure 1 is incrementally built, using the data that progressively become available, and then it is used as a *Hub* in an incrementally built *Hub-and-Spoke* architecture. The idea is that new spokes are added when there is a need of a new service which cannot be provided by the existing spokes. 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[15]. In our architecture, the Hub collects data extracted from various data sources (Extract), encodes data according to a uniform RDF-like model and ontology-based terminology (Translate) and creates a KG through a semantic data integration framework (Load). Through application-specific Application Programming Interfaces (APIs) offered by the Hub, a number of Spokes get centralized access to the KG. Each of them is granted access only to the portion of the KG that is strictly necessary to run the service. Overall, the Hub fulfils the following requirements.

It provides centralized access to data. The Hub, offers centralized access to the data that are 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 all the advantages that come from their vertical end-user applications. Advantages include contained costs, dedicated business processes, focused data, dedicated users and confined responsibilities. Data about all the key entities that are necessary to support centralized services is copied in the Hub by means of Extract, Transform and Load (ETL) facilities. ETL facilities ensure that data about the same entity extracted from multiple sources is appropriately collected, transformed, merged and correlated. In particular, entity matching (see for instance [3]) and merge facilities are essential to avoid the presence of duplicates.

It supports knowledge localization in a single university, such as the customization of the data model and terminology to be used to represent information depending on the local

[15]https://www.forrester.com/report/Deliver-On-Big-Data-Potential-With-A-HubAndSpoke-Architecture/RES83303

administrative framework. To achieve this goal, a common conceptual data model and an ontology are required for universities. Their main purpose is to provide a common core of entity types, properties and terminology in multiple languages necessary to fulfil the envisioned services within a university and to favour interoperability among universities. At the same time, the different institutional needs of universities across the globe demand for the capability of the system to support their customization and extension as required by the centralized services of a certain university.

It supports the development of centralized services. The Hub offers APIs to provide access to the KG. APIs support the development of university services on the Spokes such that they can all query the Hub and exploit the same content.

We provide below some examples of services that we have envisioned and that we are progressively developing.

Discovery services supporting browsing and search. Through them, users can issue expressive queries seeking any entity based on their properties [4]. For example, in the university scenario, users may want to search for: (a) papers written by a certain author with specific access rights; (b) courses taught by a given professor; (c) people who both teach at least a course and lead a research project on a given subject; (d) the top 10 most productive researchers in terms of funded projects. Dedicated knowledge browsers can be developed to support users in searching and browsing data in tabular, hierarchical and other visualization modes.

Communication services conveying institutional information to university stakeholders. They exploit knowledge content to offer innovative ways to present institutional information to different actors [1]. They play a crucial role in that these services support the capability of a university to present uniformly and consistently information across different institutional information channels [5]. For instance, the university may want to publish the same information on the main institutional website of the university, on the website of a specific department or of a specific professor, on the students' mobile app and on social media. Within the different Spokes, transformation procedures may take care of selecting data from the Hub and adapting it in terms of schema, language, terminology and granularity according to the different purpose and audience of the various channels.

Predictive & data analytics services supporting decision-making processes [6]. They include machine-learning applications to predict trends, and institutional dashboards that allow the governance to explore and discover correlations between data. For instance, in the university scenario, the governance may want to have a look at: (a) the trend of publications over the past 5 years within each academic department to decide whether it is necessary to provide incentives to departments in order to further improve productivity; (b) the percentage of publications in open access, to decide whether appropriate campaigns should be launched to promote a change in the publishing culture; (c) the percentage of funded projects w.r.t. those submitted to the various funding agencies, to decide whether it is necessary to help researchers write better project proposals.

Interoperability services supporting data exchange. These services support the mapping and import/export of data from/to existing standards [1]. They may also offer the capability to answer queries across multiple universities (for instance, to support students in the search for educational opportunities) and to share data with other stakeholders (for instance, with companies or with the government). A typical example of service of this kind is the centralized publication of institutional Linked Open Data [7] such that other universities and research institutes can freely re-use it. Such service should support the conversion and publication of the data in an appropriate common standard model (see for instance the VIVO model [2], designed for universities) and syntax (e.g. SKOS, RDF or JSON) linked with standard vocabularies. The publication should take place in compliance with the institutional and national regulations, especially those on data protection and intellectual property.

III. The methodology

The methodology accounts for multiple layers of diversity [8] [9] , taking inspiration from [10], and by implementing the ideas and requirements presented in [1]. The latter is based on notions and terminology familiar to the Library Science community, the first community dealing with the problem of uniformly archiving and indexing creative works. It defines an iterative process composed of sequential steps which are followed every time a new service needs to be designed and developed. Let us analyse these steps.

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

Step 2. Knowledge localization of the core ontology. The starting point is the reference data model, formalized as a core ontology (see Section IV), providing the schema which is enforced to store the data in the Hub in the form of a KG. The data model is constituted by entity types and properties necessary to describe typical key entities of universities such as people, courses, publications, dissertations and research projects. It should include identifiers, i.e. those properties necessary to identify univocally an entity of a certain type such that entity matchers can work properly [11]. In this step, the data model is extended incrementally with entity types and properties that are necessary to support the new service. [12] presents a methodology that can be followed to design the data model based on a set of user queries.

Step 3. Language localization of the core ontology. The starting point is the core ontology already available in all the desired target languages (see Section IV). For instance, it should provide the terminology necessary to describe the various roles of people (i.e. full professor, associate professor, researcher), the various kinds of publications (i.e. journal article, conference paper), the statuses of a research project (i.e. submitted, approved, funded). With localization, the core ontology is extended incrementally with concepts, relations

between them and labels in multiple languages, according to what is necessary to support the new service. See [13] for a methodology that can be followed to construct an ontology in a specific domain. A non trivial issue to be solved in this step is that of handling *lexical gaps* [14], i.e. concepts which do not have a precise translation in the target language. This fact happens quite frequently because of the different local organizations of universities.

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

Step 5. Building the KG. ETL facilities are implemented in order to Extract and Translate data into the localized data model and ontology, and to Load them in the Hub. Mechanisms to solve 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 [15]. Overlaps are handled through entity matching and merging. 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. once a day). An example of a case in which human intervention is required is to fix mistakes in the data (e.g. misspellings) or accommodate for missing terms in the ontology (thus requiring an extension of the ontology). Fixes are recorded and applied automatically in the next runnings [10].

Step 6. Implementing the service. The service is implemented and deployed by accessing the KG data via APIs.

The proposed methodology has a series of advantages.

It is scalable. The methodology ensures that the whole system infrastructure is incrementally extended and adapted to support new services – each of them being served by a different Spoke - as soon as more user needs arise. Each Spoke focuses on a subset of the data in the Hub. Given that localization leads to an extension of the data model and the ontology (in concepts, semantic relations and language), the adaptation guarantees that supported services continue to function as expected, with no need for modifications.

It is cost-effective. The data model focuses only on what is strictly necessary to accommodate for service requirements, constraints and functionalities. An example of (legal) constraint is data protection, as required by the General Data Protection Regulation (GDPR)[16], that is at the basis of privacy-by-design systems [16].

It allows for generality [17], [18]. The data model makes explicit the implicit assumptions of individual data sources. In other words, some information implicit in the legacy systems needs to be reconstructed and explicitly represented in the KG. This operation is particularly important for interoperability services. For instance, implicit assumptions may include the fact that professors are people and that all people are affiliated with the local University.

[16]https://gdpr.eu

IV. The core ontology

The core ontology, i.e., the schema of the reference KG, needs to be general enough to take into account the heterogeneity of data within a single site and across universities. The approach used in the design of the core ontology is stratified across four layers accounting for (i) concepts, (ii) language, (iii) schema and (iv) data. Each layer is built on previous ones. See [19] [8] for details. A language-agnostic conceptual layer represents knowledge with concepts and semantic relations between them. This is the layer where we abstract away the heterogeneity of languages. The language layer accommodates multiple languages for applications running within a university, (e.g., when the same local service targets users of different countries) and across universities (i.e. when a service reuses data from multiple universities). Schemas provide the reference models for representing real-world entities and their properties. They are written in terms of concepts and *not* of language-aware words; they are therefore language agnostic and apply to all languages. The data layer is where real-world instances are represented and is (mostly) language-agnostic as well.

The core ontology crucially exploits the *UKC* (for *Universal Knowledge Core*), a Wordnet-like [20] multilingual lexical resource. The UKC covers the first two layers and enables the development of the third [14], [21]. We summarize below the features which are key to this work. The idea is to tackle and develop the four layers mentioned above separately. Within the UKC, the Concept Core (CC) module provides language-agnostic concepts and semantic relations between such concepts; the Language Core (LC) module allows the definition of separate languages (e.g. English, Italian, Mongolian) as a set of lexical items connected with concepts as well as lexical relations between them. On top of the UKC, the *EType Core (ETC)* module allows for defining a set of real-world schemas, i.e., entity types and their properties, arranged hierarchically. Then finally, the *Entitybase* module (EB) allows representing instances which populate the schema generated by the ETC. Additionally, the UKC is natively integrated into NLP tools to be used in the process of integration of data, possibly from different languages. [22] provides a detailed, step-wise description of how this process works, applied to the health domain. [23] describes an early version of how the multilingual development of the UKC is performed.

Figure 2 provides a simplified view of the reference KG schema (see Step 2 and Step 3 in Section III).[17] It consists

[17]In this figure `NLString` and `SString` are two datatypes, supported by the platform we employed, which are variations of the RDF datatype `string`. The first encodes the language in which a string is written, the second stores the language agnostic concept level representation of a string, which is obtained by running NLP on the string itself [24].

of core Entity Types (classes) and a bunch of object and data properties, defined in the ETC. *Entity* is the root of our Entity Types hierarchy and represents named entities. We specialize both Entity Types and properties. For instance, *Paper*, *Book* and *Collection* specialize *Publication*, that in turn is a specialization of *Creative Work*. For properties, *Person name* specializes *Name*, *Birth date* specializes *Start*, *ORCID* specializes *Identifier* and so on. As example of relation, *Affiliation* specializes *part-of*. All the terms used in Entity Types and their properties (names and values) are mapped to corresponding concepts in the CC, and concepts are mapped to lexical items in multiple languages in the LC. This ontology is RDF complaint.

Figure 3 provides an example of KG constructed by following the data model and the ontology (see Step 5 in Section III). In the Data Layer, stored in the EB module, it shows a *Publication* authored by *Alice*, *Bob*, and *Cho* and a *Dissertation* authored by *Dan* and *Eve*. Connections between an Entity Type and the corresponding entity are exemplified by icons and colours.

The core ontology is compliant with VIVO. VIVO uses 15 existing ontologies and it further specializes their classes. In terms of coverage, given that our initial scope is narrower, our core ontology is currently smaller than VIVO. For instance, it does not include the classes that VIVO uses to represent education training and teaching process, scholarly activities, lab equipment, and materials. Still, there are some differences, especially in representing creative works and roles. One advantage of the core ontology w.r.t. VIVO is that, despite the fact that they both provide support for localized KGs, in our framework this job is modular because designers can work separately on the language, concepts, schema and data layers. In addition, while VIVO natively supports only the English language, the UKC already supports 1000+ languages [25].

V. Use cases

In this section, we present the two uses cases. For each of them we first describe the services implemented so far and then how the methodology has been applied.

A. University of Trento

The first version of the Hub-and-Spoke architecture was developed at the University of Trento in 2015 [5] . The goal was and still is to enhance the information assets of the University of Trento through the adoption of data management strategies that ensure their quality and encourage data reuse.

1) Services: Four services have been implemented so far as follows.

Institutional portal. It is a communication service that offers a unified view of the University members, academic departments, governing bodies and administrative units[18]. Members include academic staff (professors, researchers, PhD students), administrative and technical staff, and university executives. The service provides contact information (email addresses, phone numbers, addresses), CVs, list of publications, courses, projects, master and PhD theses. The portal is in English and Italian and it is visited by around half a million users per year.

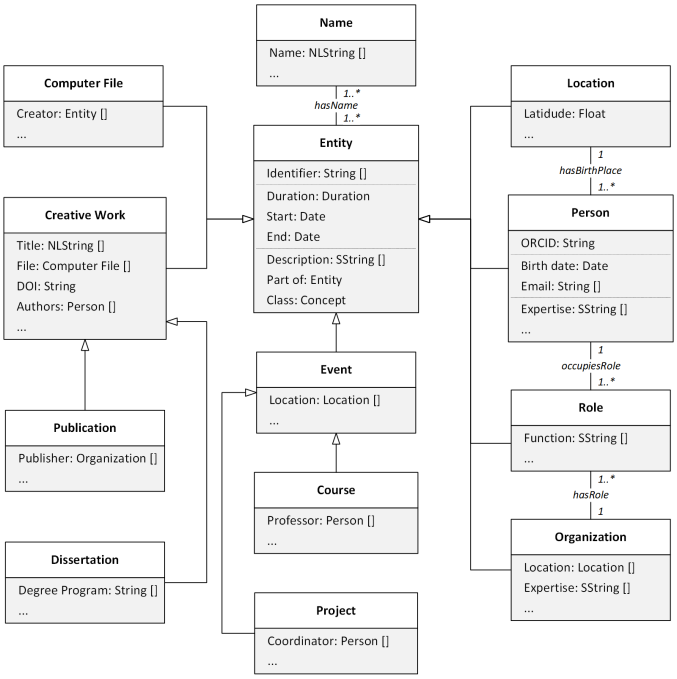

Fig. 2. Partial view of the University data model and ontology

Institutional dashboard. It is a data analytics service providing insights about the quality of research conducted by the faculty members. It helps decision makers with statistics and interactive graphs that are useful to the University governance to examine trends, strengths, and points of improvement. Access is reserved to University members only.

Open Data publication. It is an interoperability service that supports the publication of Open Data on the regional[19], national[20] and European[21] data portals. Thus, the University complies with national guidelines about sharing public sector information. It is important to notice that the published data are of top quality (uniform schema and terminology, offered in English and Italian, with entities in different datasets that link to each other via unique identifiers) and that this step had no cost, only that of extracting the data from the KG and publishing them in the appropriate format.

University Mobile App. A second communication service has been developed by the IT staff of the University. It is a Mobile App[22] for students that partially uses data of the Hub through dedicated APIs.

2) The development process: To collect requirements (Step 1 of the methodology), we interviewed a large number of potential users. For the portal, we interviewed students and

[18]https://webapps.unitn.it/du/en

[19]https://dati.trentino.it/organization/universita-di-trento

[20]See for instance: https://www.dati.gov.it/view-dataset/dataset?id=6f65d051-e48e-475b-a22b-6b0c1267cf96

[21]See for instance: https://data.europa.eu/data/datasets/theses-of-the-university-of-trento?locale=en

[22]https://unitrento.app

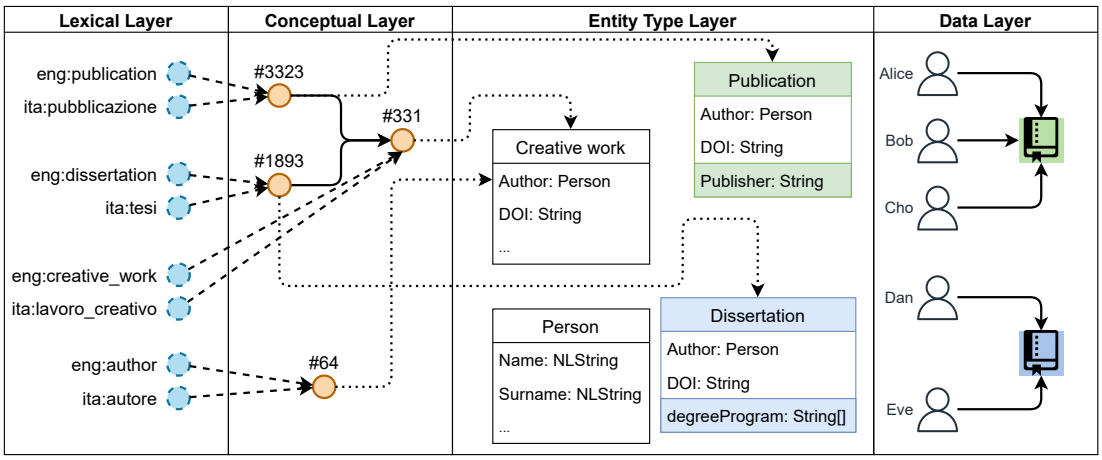

Fig. 3. A simplified view of the layered representation of a KG

technical, administrative and academic staff. For the dashboard, we interviewed the heads of our academic departments and the director general. In addition, we reviewed the portals of the top 30 universities according to the Times Higher Education. The staff of the Communication division designed the mock-ups of the services. We also conducted a user study to validate and refine the mockups. The localization of the data model (Step 2) was done by inferring the additional properties to be added to the core model from the mockups of the services. The localization of the ontology (Step 3) consisted in adding more concepts and corresponding labels in English and Italian. Concepts were required for entity properties and corresponding values. They include, for example:

- 42 types of roles played by people (full professor, associate professor, dean, technical staff, PhD student, ...) that were inferred from around 250 different labels used in IT systems to denote similar roles (thus reducing diversity);
- 14 types of organizations (academic department, directorate, office, governing body, ...) that were inferred from around 100 different labels used in IT systems to denote similar organization types;
- 22 types of publications (journal paper, conference paper, book, chapter of a book, ...);
- 57 values for the language of a publication (English, Italian, Spanish, ...);
- 7 values for the status of a project (submitted, rejected, funded, ...);

Data are extracted from 7 different IT systems (Step 4). Two of them were designed to accommodate data not yet available in legacy systems (Step 4, case c). The first one accounts for names and descriptions of organizations in English (initially, only names in Italian were available), and maps the original terms in the data sources with the concepts in the ontology. The second system allows people to provide their photos, CVs, notices, office hours and thesis proposals.

ETL facilities (Step 5) were written to import data into the Hub from the legacy systems. In the initial version of the Hub, the KG contained about 1 million entities. Later in the project, we discovered ways to eliminate those entities irrelevant to the services offered from the KG. Currently it contains about 250.000 entities, including people, organizations, publications, dissertations, files, patents, courses, and projects. Services were developed (step 6) mainly in Java and AngularJS and use Elasticsearch[23] for data indexing.

B. National University of Mongolia

A feasibility study for digital services at the National University of Mongolia was performed in April 2020 with inspiration from the successful implementation in Trento [5]. It includes a shallow assessment of functionalities in legacy systems w.r.t. data exchange, architecture, existing and missing data, and possible expected outcomes. First, we constructed a small KG based on open data published by the National University of Mongolia[24] as data infrastructure. Secondly, we developed a minimal viable product, a simple chatbot, that answers questions about faculties, courses, course schedules they teach, etc. It works in two information channels: web chat and Microsoft Teams. In October 2020, the National University of Mongolia pilot project was officially launched with a promising proposal and has been delivered in September 2022.

1) Services: Two centralized services, a chatbot and a GraphQL API, have been developed.

Chatbot. The chatbot is an NLU-based question-answering system that answers questions centered on faculties and staff in the University. Users can ask questions in free text in Mongolian. The chatbot replies with relevant information from the KG by using identified intents and named entities in the questions. The chatbot organizes a set of predefined question groups for machine learning models. The idea beyond the chatbot is that a unified conversational user interface could be expanded for further services. In other words, this bot

[23]The Elasticsearch website: https://www.elastic.co/products/elasticsearch
[24]http://data.num.edu.mn

framework could be a digital assistant capable of including all the services around the University.

GraphQL API service. It is an interface to access the KG, and it offers fetching complex and user-defined data with a single API call. This API fits complex systems like bot frameworks and other microservices.

2) The development process: In Step 1 of the methodology, we took a small survey of faculties and students and asked about their interest in the information, frequency, and information channels at their convenience. We also interviewed administrative staff w.r.t. what type of services are needed. As a result, potential microservices were identified, and one of them was centered on promoting academic staff and the usage of resources like courses and rooms. In this phase, we also conducted a quality assessment of data in legacy systems. The data quality assessment was based on four main criteria [26]: accuracy, completeness, consistency, and timeliness. The overall data quality score was 0.84.

In Step 2, we identified additional entity types to be addes to the core model and their corresponding properties needed and qualified for the chatbot service. In Step 3, the core ontology was expanded by adding more concepts and labels in English that are required for the entity types and properties. We added 8 new classes and 20 properties. In Step 4, we identified relevant data from three legacy IT systems, i.e. the Student Information, Research Management and Project Management systems. They include data about organizations, articles, professors, journals, projects, courses, and conferences. In Step 5, several simple web APIs were developed on top of the legacy systems for extracting related data. With KarmaLinker, an extended version of Karma[25] data integration tool [10], we developed data integration models using extracted data and the localized ontology. We created a small ETL microservice in NodeJS that runs data integration models and publishes Resource Description Framework (RDF) data to a triple store that provides a SPARQL endpoint. The ETL tool imports RDF into the Neo4J graph database management system from this triple store and provides data through the API.

In Step 6, the chatbot service was realized. It consists of a bot framework developed on Rasa[26] open source and an action server written in Python. The bot framework is featured with NLU capabilities and is implemented in the bot application logic. The action server connects the bot framework, the information channels, and the GraphQL API. At the time of writing, the chatbot service supports only the Microsoft Bot Framework for the Microsoft Teams channel.

VI. Challenges

During the development, we had to face challenges that typically arise when dealing with new IT services [27]. Organizational challenges pertain to the obstacles that need to be overcome in order to move from consolidated practices to new ones. Technical challenges relate to the difficulties concerned with the identification or the creation of appropriate technologies. Conceptual challenges relate to the difficulty of identifying and adopting the proper standards. Legal and security challenges include dealing with Intellectual Property Rights (IPR), licensing, security and privacy. Last but not least, there are all the challenges related to the end-users. Let us see how these challenges manifested themselves in these projects.

Organizational challenges. We had to convince the governance of the universities to invest in the two projects. This task was achieved by providing concrete examples of problems that need to be solved and that we actually tackled through proof of concept projects that lasted 1 year for University of Trento and 8 months for National University of Mongolia. During this period, a sample of the data sources was integrated by means of an initial simplified ETL framework. Demos of envisioned services were developed and presented to institutional bodies of the two universities. The University of Trento project team closely collaborated with its Legal, IT, Library and Communication departments. The National University of Mongolia team worked with its IT and Research Department. We believe that the achievements we have been able to accomplish would not have been possible without such tight collaboration. Both universities set up a clear project plan with defined tasks, deadlines and responsibilities. University of Trento also created a University Committee, with a member from each academic and administrative department, that had an important role in establishing the project goals, in collecting service requirements and favoring the adoption of the services.

Technological challenges. These challenges vary in technology and resources deployed in the two universities. As described above, the first version of the University of Trento's Hub is an instance of the Semantic Web technology developed by researchers at University of Trento, so called SWEB. SWEB offered all the necessary functionalities and user interfaces to operate on the UKC, the ETC and EB modules (see Section IV). Yet, at the initial stage of development of the project, SWEB soon showed some limitations in terms of performance. Therefore, University of Trento had to distribute the workload on 3 virtual machines to make sure that the ETL process was fast enough to guarantee periodic updates. University of Trento progressively adopted several tricks to tune and improve performance. Currently, the entire infrastructure is under refactoring to further improve performance and reduce the number of technologies used both on the ETL and on the service side. On the other hand, National University of Mongolia decided for a more standard OWL implementation of the ontology, used to integrate data to build the KG, with no particular difficulty experienced.

Conceptual challenges. The project teams developed the core data model and ontology by extending those previously employed in similar research projects carried out at University of Trento. Such development took significant time.

Legal and security challenges. We adopted privacy-by-design principles, and in particular, the strategies proposed by Hoepman [16]. Such strategies have been suggested by

[25]https://usc-isi-i2.github.io/karma/

[26]https://rasa.com

the European Data Protection Supervisor (EDPS) as a good example of an approach that can be followed for identifying measures to implement privacy requirements. In practice, this means that privacy had to be considered a fundamental requirement since the design of the entire system infrastructure. University of Trento started with a Data Protection Impact Assessment (DPIA) document describing risks and the strategies employed to reduce them, the characteristics of the architecture and the services to be implemented. University of Trento had to obtain the formal approval of the Data Protection Officer (DPO) of the University. Each service is accompanied by its own "Cookie policy and information on the processing of personal data". In addition, University of Trento had to comply with the programming and security standards of the University and the national guidelines of the AgID. In particular, University of Trento had to guarantee secure access to only authorized users to the Hub and the system administration services. For the Open Data service, University of Trento had to promote the adoption of a new dedicated regulation ("Regulations on access to University documents and data") and to obtain the formal approval of the Academic Senate and the Rector. In terms of Intellectual Property Rights (IPR), University of Trento decided to promote and support the download of Open Access publications through the communication services. Each publication is accompanied with a clear indication of its license.

User-related challenges. We believe that one of the major risks to be managed is failing to meet user expectations in terms of functionalities offered and time of delivery. We mitigated this risk by ensuring proper and constant communication with them. We involved users in all stages of the work. At University of Trento we run 3 different usability studies that allowed us to refine the services.

VII. Conclusions

In this paper, we presented a methodology and a infrastructure that is being used to achieve both internal and external semantic interoperability in universities. The two projects carried out in Italy and Mongolia show the potential of the proposed solution. Both universities plan to continue developing new services. Furthermore, plans of adoption by other universities are underway. After this initial validation phase, the technology will be made available open source.

References

[1] V. Maltese and F. Giunchiglia, "Foundations of digital universities," *Cataloging & Classification Quarterly*, pp. 1–25, 2016.

[2] K. Börner, M. Conlon, J. Corson-Rikert, and Y. Ding, "Vivo: A semantic approach to scholarly networking and discovery," *Synthesis lectures on the Semantic Web: theory and technology*, vol. 7, no. 1, pp. 1–178, 2012.

[3] J. Wang, G. Li, J. X. Yu, and J. Feng, "Entity matching: How similar is similar," *Proceedings of the VLDB Endowment*, vol. 4, no. 10, pp. 622–633, 2011.

[4] F. Giunchiglia, B. Dutta, and V. Maltese, "From knowledge organization to knowledge representation," *KO KNOWLEDGE ORGANIZATION*, vol. 41, no. 1, pp. 44–56, 2014.

[5] V. Maltese, "Digital transformation challenges for universities: Ensuring information consistency across digital services," *Cataloging & Classification Quarterly*, vol. 56, no. 7, pp. 592–606, 2018.

[6] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management," pp. 77–84, 2013.

[7] E. Tran and G. Scholtes, "Open data literature review," *Barkeley School of Law, University of California*, 2015.

[8] F. Giunchiglia, A. Zamboni, M. Bagchi, and S. Bocca, "Stratified data integration," *arXiv preprint arXiv:2105.09432*, 2021.

[9] V. Maltese, F. Giunchiglia, K. Denecke, P. Lewis, C. Wallner, A. Baldry, and D. Madalli, "On the interdisciplinary foundations of diversity," in *The First International Workshop on Living Web: Making Web Diversity a true asset (at the ISWC 2009)*, Washington DC, USA, 25 October 2009.

[10] F. Giunchiglia, S. Bocca, M. Fumagalli, M. Bagchi, and A. Zamboni, "iTelos–purpose driven knowledge graph generation," *arXiv preprint arXiv:2105.09418*, 2021.

[11] P. Bouquet, H. Stoermer, and X. Liu, "Okkam4p: A protégé plugin for supporting the re-use of globally unique identifiers for individuals in owl/rdf knowledge bases." in *SWAP*. Citeseer, 2007.

[12] U. Chatterjee, F. Giunchiglia, D. P. Madalli, and V. Maltese, "Modeling recipes for online search," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2016, pp. 625–642.

[13] F. Giunchiglia, B. Dutta, V. Maltese, and F. Farazi, "A facet-based methodology for the construction of a large-scale geospatial ontology," *Journal on data semantics*, vol. 1, no. 1, pp. 57–73, 2012.

[14] F. Giunchiglia, K. Batsuren, and A. Freihat, "One world - seven thousand languages," in *19th International Conference on Computational Linguistics and Intelligent Text Processing*, Hanoi, Vietnam, 2018.

[15] X. L. Dong and F. Naumann, "Data fusion: resolving data conflicts for integration," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1654–1655, 2009.

[16] J.-H. Hoepman, "Privacy design strategies," in *IFIP International Information Security Conference*. Springer, 2014, pp. 446–459.

[17] J. McCarthy, "Generality in artificial intelligence," *Communications of the ACM*, vol. 30, no. 12, pp. 1030–1035, 1987.

[18] P. Bouquet and F. Giunchiglia, "Reasoning about theory adequacy. a new solution to the qualification problem," *Fundamenta Informaticae*, vol. 23, no. 2, 3, 4, pp. 247–262, 1995.

[19] F. Giunchiglia, V. Maltese, and B. Dutta, "Domains and context: first steps towards managing diversity in knowledge," *Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web*, pp. 53–63, 2012.

[20] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[21] F. Giunchiglia, K. Batsuren, and G. Bella, "Understanding and exploiting language diversity," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 2017, pp. 4009–4017.

[22] G. Bella, L. Elliot, S. Das, S. Pavis, E. Turra, D. Robertson, and F. Giunchiglia, "Cross-border medical research using multi-layered and distributed knowledge," in *Proceedings of Prestigious Applications of Intelligent Systems (PAIS@ECAI)*, Santiago de Compostela, Spain, 2020.

[23] A. Tawfik, F. Giunchiglia, and V. Maltese, "A collaborative platform for multilingual ontology development," *World Academy of Science, Engineering and Technology*, vol. 8, no. 12, 2014.

[24] G. Bella, L. Elliot, S. Das, S. Pavis, E. Turra, D. Robertson, and F. Giunchiglia, "Cross-border medical research using multi-layered and distributed knowledge." *KI-Kunstliche Intelligenz*, 2020.

[25] G. Bella, E. Byambadorj, Y. Chandrashekar, K. Batsuren, D. A. Cheema, and F. Giunchiglia, "Language diversity: Visible to humans, exploitable by machines," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, 2022.

[26] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, jul 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541883

[27] V. Maltese and F. Giunchiglia, "Search and analytics challenges in digital libraries and archives," *ACM Journal of Data and Information Quality*, 2016.