computers

MDPI

Article

Expressing the Tacit Knowledge of a Digital Library System as Linked Data

Angela Di Iorio * and Marco Schaerf

DIAG—Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome, Via Ariosto, 25, 00185 Rome, Italy; marco.schaerf@uniroma1.it
* Correspondence: angela.diiorio@uniroma1.it

Received: 7 May 2019; Accepted: 11 June 2019; Published: 15 June 2019

check for updates

Abstract: Library organizations have enthusiastically undertaken semantic web initiatives and in particular the data publishing as linked data. Nevertheless, different surveys report the experimental nature of initiatives and the consumer difficulty in re-using data. These barriers are a hindrance for using linked datasets, as an infrastructure that enhances the library and related information services. This paper presents an approach for encoding, as a Linked Vocabulary, the "tacit" knowledge of the information system that manages the data source. The objective is the improvement of the interpretation process of the linked data meaning of published datasets. We analyzed a digital library system, as a case study, for prototyping the "semantic data management" method, where data and its knowledge are natively managed, taking into account the linked data pillars. The ultimate objective of the semantic data management is to curate the correct consumers' interpretation of data, and to facilitate the proper re-use. The prototype defines the ontological entities representing the knowledge, of the digital library system, that is not stored in the data source, nor in the existing ontologies related to the system's semantics. Thus we present the local ontology and its matching with existing ontologies, Preservation Metadata Implementation Strategies (PREMIS) and Metadata Objects Description Schema (MODS), and we discuss linked data triples prototyped from the legacy relational database, by using the local ontology. We show how the semantic data management, can deal with the inconsistency of system data, and we conclude that a specific change in the system developer mindset, it is necessary for extracting and "codifying" the tacit knowledge, which is necessary to improve the data interpretation process.

Keywords: linked data; tacit knowledge; ontology management

1. Introduction

This paper is an extension of the paper [1] presented at the REMS 2018 Multidisciplinary Symposium on Computer Science and ICT in Stavropol, Russia, at the North–Caucasian Federal University. The ontologies used for exhibiting the linked dataset were already presented in the conference version of the paper. The new contribution in this paper is refining the semantic data management method, adopted for extracting the "tacit" knowledge of a case study system, and prototyping the generation of the corresponding linked dataset, as a result of the semantic web technology (SemWebTech) implementation, in the data management practices of an organization (ORG).

A digital library (DigLib) is a long-standing ORG, managing multi-media objects, from their acquisition, through their entire digital life-cycle. Usually DigLib systems manage multi-media objects, on a long term perspective basis, as such, data describing the multi-media maintenance process, is ever-growing.

Thus, the implementation of SemWebTech in an existing information system (InfSys) like a legacy DigLib system, should be considered an essential evolution, because it allows us to deal with the pervasiveness of technologies, and big data challenges, related to the increase of systems' complexity.

The data management practices of a DigLib system are indeed challenged by the need of maintaining the accessibility to the multi-media objects in the long term, and by the evolution of the holding InfSys, as a set of humans, technologies, data, information and knowledge. The linked data (LD) initiative aims to build an interconnected database where semantics and data can be used globally (openly or not) allowing anybody to use it, as an infrastructure. LD is essential for achieving the SemWeb vision, and it is supposed to be re-used by consumers (humans and machines) over web protocols. Actually this is still difficult to flawlessly achieve. Many of the pioneering projects have produced datasets characterized by a quality level, which is not sufficient to facilitate data re-use [2], because the interpretation process is time-consuming and misunderstanding-prone, unless human consumers are already experts of the knowledge domain.

In a library domain, most of the initiatives are primarily experimental in nature, as reported by Smith [3], and Tosaka and Park [4] still report the "significant problem of the absence of comprehensive data, that could be used to guide improvements in continuing education" for the library community.

In this paper, we present our approach for generating LD from the relational database of a DigLib system case study, and we show how we have addressed the capture of the data context, according to the Linked Data Best Practices, published by the World Wide Web Consortium [5] and related glossary [6] (Modeling process: https://www.w3.org/TR/ld-glossary/#modeling-process):

> [...] capture the context of data [...] high quality of Linked Data is obtained since capturing organizational knowledge about the meaning of the data within the Resource Description Framework (RDF) [7] data model means the data is more likely to be reused correctly. Well defined context ensures better understanding, proper reuse, and is critical when establishing linkages to other data sets.

Our approach is analyzing the organizational knowledge, as the key component driving the interpretation of data, and in particular the "tacit" knowledge that is not already made explicit by the existing SemWeb ontologies (i.e., LD vocabularies (LOV) [8]), nor stored as facts in the relational database, the data source.

We analyzed a relational database, used by a DigLib system of the Sapienza University. We found that, even semantically defined by existing ontologies regarding to well known DigLib metadata standards (adopetd by the system), data is managed by relying on a "tacit" [9] ORG knowledge. This knowledge should be made explicit, as an ontology, in order to support the data interpretation of consumers, and to improve the conveyance of the data meaning.

The capture of "tacit" ORG knowledge is addressed by managing data, and its knowledge, taking into account the LD consumers' perspective, by driving data management toward the "semantic data management". "Semantic data management practices" require us to establish, in the holding ORG, a mindset specifically oriented toward the pillar elements of the LD, like the uniform resource identifiers (URIs), and the vocabularies (controlled term lists, thesauri, ontologies, etc.), exhibited as a linked open vocabulary (LOV) [8]. Semantic data management practices imply (1) to consider InfSyss' knowledge as a data in the data management practices, (2) to capture cultural semantic elements that have influence on the InfSys, (3) to curate, at best, the semantics that can better express the data meaning, and support the correct interpretation of LD consumers.

We have already presented, in the REMS 2018 conference [1], two local ontologies obtained by analyzing the DigLib system's relational database. Proposed ontologies capture the underlying "tacit" knowledge, of the DigLib system and are connected by ontological matching with existing ontologies, from wider knowledge domains. Thus, the matching, between local ontologies and existing ontologies, drives the generation of linked datasets, whose meaning is supported, not only by existing ontology, but also by the tacit knowledge expressing the ORG knowledge, mentioned by the LD best practices and considered essential for the correct interpretation of generated LD. In this paper, we also present our approach for preparing the generation process of linked dataset using local ontologies. We show the prototyped RDF [10] triples, expressing the facts stored into relational database used by the DigLib system case study.

The remainder of this paper is structured as follows. Section 2 reports the problem statement about the LD implementation in ORGs managing DigLibs. Section 3 provides an explanation of the semantic data management. Section 4 focuses on the "tacit" knowledge types. Section 5 overviews the DigLib system case study. Section 6 describes the method for detecting knowledge types from the DigLib system. Section 7 describes how we have applied the LD principles for codifying the local ontology as a linked data vocabulary and presents a graphical representation of it. Section 8, shows how we have managed data and vocabularies for preparing the LD production process, by adopting a "semantically oriented" approach, toward detected matching ontologies and LD principles, which drives a deeper analysis of the data source. Section 9 presents resulting named individuals computed from the data source of the case study. Section 10 highlights limitations, draws conclusions and presents the future developments.

2. Problem Statement and Contributions

The automation of semantics, as a piece of knowledge about data, is not a straightforward task. SemWebTechs, have been developed for achieving this goal, nevertheless their implementation in legacy InfSyss is a challenging task for data managers. Related literature reports plenty of approaches, providing mapping between relational databases and ontologies, foremost classified by Sahoo et al. [11] in: (a) automatic mapping generation and (b) domain semantic-driven mapping generation. In addition, many tools and techniques, derived from the Bizer and Cyganiak approach [12], can be used for extracting or linking semantics to relational data, as well as for enriching LD triples with semantic connections by using automatic annotation tools (i.e., in Beneventano [13]). Our approach extends the focus on the facts, related to an InfSys, that are not stored into the relational database (RDB), and points to capture and codify knowledge that is "hidden" or "given" by the InfSys organizational culture, and that influences the data management.

Alavi et al. [14] highlighted the increasing interest of InfSys researchers around the knowledge "as a significant organizational resource" and traced a review of knowledge management systems developed until that period. Knowledge and its management is a complex and multi-faceted concept, and the objective of knowledge management systems is the creation, transfer and application of knowledge in organizations. Related literature embraces different fields and the information technology plays an important role in supporting the process of the knowledge management. Furthermore the problem of capturing knowledge in software systems development and maintenance is a long-standing problem, and in particular to turn "tacit" [15] knowledge into a "describable" piece of data, is challenging. Exhibiting the ORG knowledge by means of a SemWebTech, like LD, is even more challenging because the knowledge "given" in the circumscribed area of the native organization, cannot be easily interpreted by consumers (people and machines). Thus, it is necessary to provide the wise organizational context for allowing LD consumers to correctly interpret data, and for improving the re-use process.

We agree with de Vasconcelos et al. [16] that, the knowledge management practices in software engineering would improve both software development, and more particularly software maintenance, and that a lot of knowledge remains in the individuals' mind, because capturing knowledge for later reuse is likely to be seen as a low priority. This fact creates a potential organizational knowledge gap, that reverberates into the data management.

We believe that, by adopting semantic data management practices, the tacit knowledge can be captured, codified and managed beside data. The method applied to the DigLib case study, and prototyped LD triples, contribute to provide data with its own organizational context, and then to increase the LD quality for a correct interpretation and a proper re-use. The perspective of end-users as LD consumers, should be accurately considered in data management practices of ORG producing and publishing LD. The lack of proper semantic context hinders the understandability of exhibited LD, thus the knowledge that "remains in the individuals' mind" of developers or data managers is the point of interest, for the semantic data management practices.

3. The Semantic Data Management

The semantic data management captures semantic context of data, "codifying" (see next Section 4), and encoding it, in a SemWeb language, that can be used, not only by humans, but also by machines. Data is traditionally managed by persons responsible for the InfSys of an ORG, and their practices are influenced by their human information, implicit and explicit knowledge, wisdom [17]. In semantic data management practices, data is equipped with its comprehensive semantic context, which is managed aside data and is encoded in a machine-interpretable form (SemWebTech). The detection, capture and codification of the "tacit" (also hidden or given) knowledge is a key activity of semantic data management practices. Encoding such a knowledge in a SemWeb language supports consumers in re-using or re-managing data in a proper way, because semantics convey the knowledge necessary to understand "why data has value" and "how it was managed". The interpretation of data is facilitated for humans that might be unloaded by long and discontinuous searches of additional information in interpreting data, for machines that could re-use data with more accuracy.

The literature is rich of works generating ontologies from data and metadata of relational databases, but the underlying knowledge is not always explicit, and it is scattered into technical reports, or documentation. Software and data documentation is a daunting task, that should come up besides the system development, but it is often neglected to the point of being completely missing. Thus the process of interpreting data and the system functionality results in a time-consuming task, because its knowledge is difficult to retrieve and accessed.

The proposed semantic data management applies at the data source, and aside from the current data management practices of a legacy system. RDF is produced already provided with the source ORG context, allowing LD consumers to be aware about why data belongs to a dataset, and how data was managed in the ORG context. Indeed, the method is not limited to the data source but aims to extract knowledge from the culture of the system managers, and as such, it creates knowledge resources about the database management that are not mentioned by database metadata. In other words, the knowledge of a domain expert, that usually validates an ontology, is already provided at the source of data, and is extended to the context, within data is managed. The "tacit" knowledge, inadvertently implicit, hidden or simply given, into the data management practices of the system, is essential for enabling machines to properly interpret data, thus its capture, and its codification as a SemWeb vocabulary (i.e., LOV), is an essential part of the semantic data management. The adoption of SemWebTech in an existing data management system, implies to capture the relevant knowledge supporting the data interpretation of a consumers, thus the semantic task of a LD producer should focus on the existing "explicit" knowledge about data management, and should turn explicit the "tacit" knowledge.

4. Focus on the Tacit Knowledge

In the early nineties, Wiig [18] pointed out that the ORG knowledge is one of the most nebulous and difficult concept to define. Evans et al. [19] report that this kind of indistinct knowledge has been defined in related literature, as an ORG "asset" for its own. Boisot [20] also defines "knowledge stocks" through which a variety of value added services flow, and classifies them as abstracted principles or as codified form. We believe that, the degree, to which the knowledge assets (underlying an InfSys) can be expressed in an explicit form, depends on the mindset of the system developers and data managers. The ontologies presented in REMS 2018 [1] represent the result of the approach we have adopted for detecting, capturing and codifying the ORG knowledge of the DigLib system case study. Similarly to what is theoretically stated for the knowledge management in the ORG borders, the knowledge about data should comprise of explicit and "tacit" [9,21] knowledge, the knowledge that is still not codified.

According to the knowledge management literature, we discuss explicit knowledge classification, in relation to the SemWebTech implementation in our case study:

- "Codified" knowledge is highly refined, and formalized, to the point in which it can be written down, lowering the risk information losses [22].

  SemWeb ontology supplies a way of formalizing knowledge "stocks", and extends the re-use possibilities, not only to humans, but also to machines.
- "Encapsulated" knowledge is usually not completely codified, and it is object-embedded, since the knowledge necessary to the object design and development remains partially hidden from its users [23].

  Software artifacts and databases well represent this kind of knowledge. As we have already pointed out, many knowledge is given, missing, mis-documented or not easily understandable, due to the fast increase of systems' complexity and the fast pace of InfSys evolution. This speed entails to neglect the slow and expensive task of documenting such kind of knowledge. Semantic data management practices focus on this kind of knowledge.

Both "codified" and "encapsulated" knowledge derive from the "tacit" knowledge, that originates from thought, reflection or experience and remains resident in the human mind. Tacit knowledge provides the "grounding of meaning" and the basis for the interpretation of a tacit activity [19].

The ontology conceptualized from the DigLib system case study, and described in the following sections, is considered the "grounding of meaning" for LD to be published, and "codified" for facilitating the understandability [24] of humans and machines.

5. Digital Library System: A Case Study

This section describes the massive conversion (MassConv) system, the DigLib system case study. Figure 1 shows an abstract overview of the data management performed by the MassConv system, as a component of the Sapienza digital library (SDL) [25], an InfSys working on behalf of the Sapienza University. MassConv is a data management system based on RDB, where managed data are collected from the SDL InfSys, and used for producing digital resources (DigRes) (see Figure 2) conforming with well-established DigLib metadata standards. Data, managed by the MassConv system, describes the SDL multimedia objects.

The system was developed for managing the DigRes production workflow, based on Information integration global-as-view approach [26], where the RDB (the data source $\mathcal{S}$) contains data about digitization projects, undertaken by Sapienza University (Documentazione delle Risorse Digitali di Sapienza Digital Library, Indice dei progetti e delle iniziative di digitalizzazione, https://sbs.uniroma1.it/sapienzadl/it/index_progetto_digitalizzazione), and about the management of digitized items.

By using a local mapping $\mathcal{M}$, MassConv automatically converts managed data into XML files, conforming with global schema $\mathcal{G}$, represented by the XML schemas of the following standards:

- the metadata objects description schema (MODS) (metadata encoding transmission schema, www.loc.gov/standards/mods/), describes the intellectual contents represented by multimedia objects;
- the preservation metadata implementation strategies (PREMIS) [27] encompasses preservation metadata about multimedia objects;
- the metadata encoding and transmission standard (METS) (metadata encoding transmission standard, www.loc.gov/standards/mets/) comprehends the data for packaging descriptive and preservation metadata, and multimedia objects.

In this paper the knowledge detection of the DigLib system (see Section 6) is focused on the MODS and PREMIS metadata, that in the MassConv system are respectively managed as descriptive and preservation metadata, which is abstractly showed in the Figure 1, as an RDB supporting the MassConv workflow steps.

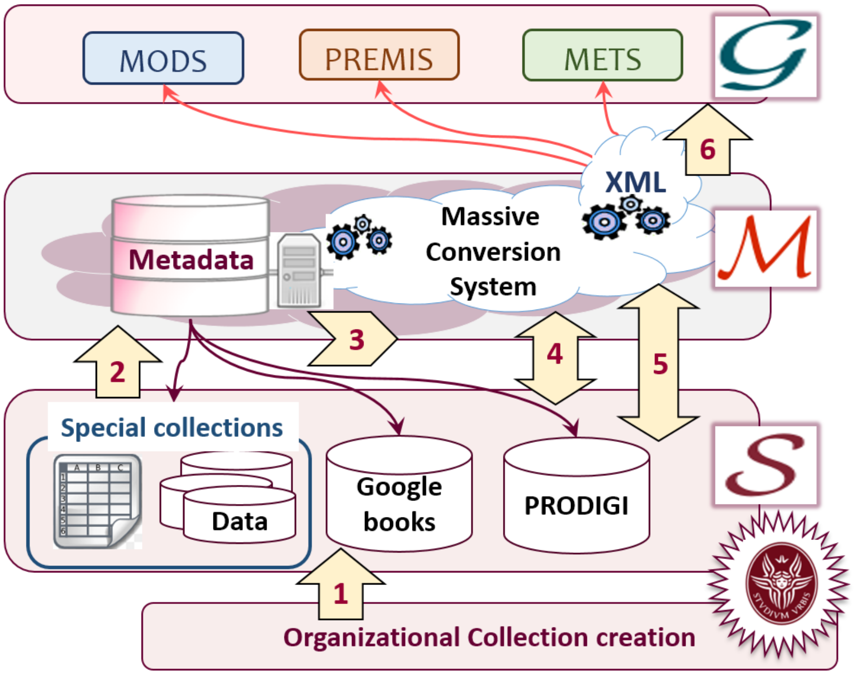

Figure 1. Abstract representation of architectural elements, and data workflow of the Sapienza digital library (SDL) massive conversion system.

MassConv workflow steps are depicted as yellow numbered arrows, on purpose we will describe performed steps in the Section 6.3, where the knowledge of the software is captured.

Produced XML files are associated, by URI reference, with the multimedia objects, collected by the SDL InfSys. Figure 2 shows the model of the SDL DigRes (structural model of a digital resource, managed by the Sapienza library system, https://sbs.uniroma1.it/data/documentation/DigitalResource). SDL DigRes is an information package (IP), composed by a set of data and multimedia object. IP is used for different functional roles, as defined by the Open Archival Information System (OAIS) [28], respectively, submission (SIP), archival (AIP) and dissemination (DIP). By the conceptual point of view a SDL DigRes is the "simplest" set of information coherently managed by the MassConv system, it describes an intellectual entity [27], collected by SDL. By the structural point of view a SDL DigRes, is at least composed by digital metadata objects (DMO), a set of data and its metadata, describing multimedia objects, that are generally defined as digital content objects (DCO). DCO is the other structural component of a DigRes. Based on the type of DCO, content models are well-defined in the MassConv system, in order to manage data and multimedia object coherently to the content type.

It is worth noting that "digital collection" is the unique content model, which connects DigRes and Sapienza ORGs, by means of a continuous arrow line, stating the mandatory connection between DigRes and the ORGs, responsible for its management.

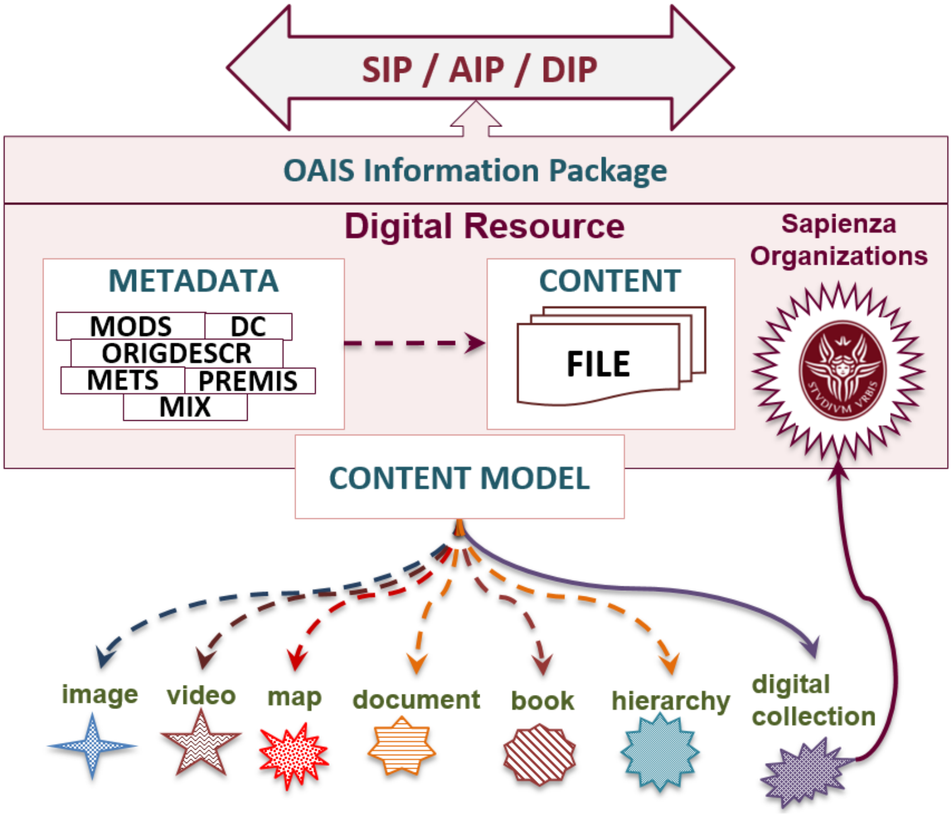

Figure 2. Structural model of a SDL digital resource, related to the different Open Archival Information System (OAIS) information package (IP) functional roles: submission, archival, and dissemination.

6. Detecting Knowledge Types in the Case Study

Figure 3 shows how we have approached the types of knowledge in the DigLib as the ORG context of the MassConv system.

1. "codified" knowledge: comprehends semantics already codified by MODS and PREMIS metadata standard (described in the Section 5) adopted by the system, and depicted as rounded boxes with a thick border. The rounded boxes, with a dashed border, represent knowledge to be codified from the "tacit" (ORG-K) and the "encapsulated" (DLsys-K).
2. "tacit" knowledge: the capture was generically focused on the DigLib available documentation and software functions. We analyzed the knowledge elements that we used for communicating between people involved in the project, and for conceiving the system. Then we have outlined, what are the knowledge elements that are not explicitly codified in the MassConv system. We observed, that elements belonging to tacit knowledge were partially documented or considered as given in the software functions. Anyhow those elements are not codified in a SemWeb language for being used by humans and machines.

3. “encapsulated” knowledge has to be extracted from the software artifacts, and used RDB.

The order established for addressing the knowledge classification process is not casual. We firstly started from (1) what is already known and codified by existing ontologies. Then we proceeded by addressing (2) what is tacitly known by SDL developers and managers, thus not codified and not encapsulated in the MassConv system. In the end, we finished with (2) what is encapsulated in software artifacts, and in the RDB, by codifying the MassConv workflow steps.

Encapsulated knowledge was addressed as the last, for reducing the risk of misappropriation [23] or codification redundancy. We distinctly represent the semantics conceptually defined, and web format agnostic, in serif font style, and the semantics defined by a web schema (usually expressed in extended markup language (XML) syntax), in courier font and prefixed by the schema name (i.e., `premis:agent`).

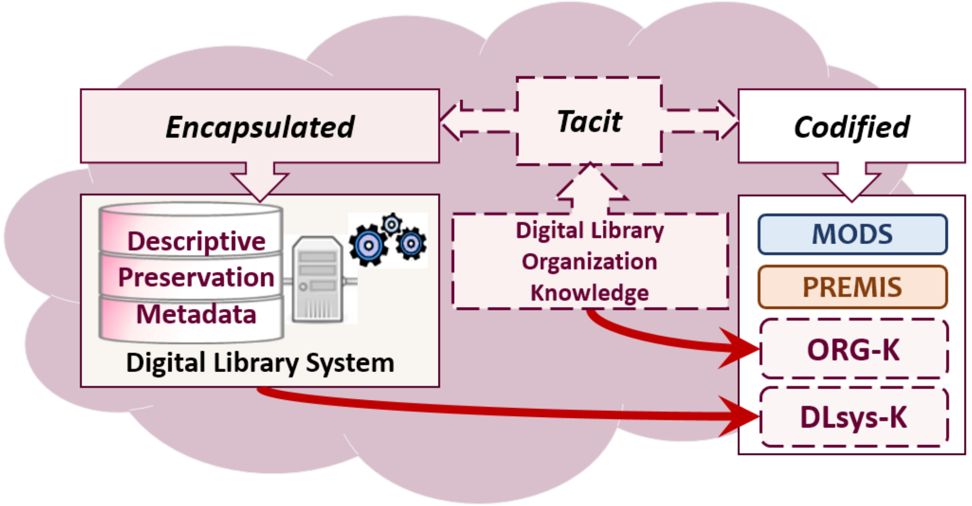

Figure 3. Detected organizational knowledge in a digital library system.

6.1. “Codified” Knowledge Capture and Enrichment

Considering the metadata semantics used by the MassConv system we briefly present the ontologies strictly representing the knowledge, “codified” by MODS and PREMIS metadata standards.

- metadata object description ontology (MODS-O) [29] develops around the main class mods:ModsResource which represents “any library-related resource—such as a book, journal article, photograph, or born-digital image—that is described by a MODS resource description”.
- PREMIS ontology (PREMIS-OWL) [30–32] models the knowledge domain of digital preservation metadata, and develops around four main classes: `premis:Object`, `premis:Event`, `premis:Agent` and `premis:Rights`.

By analyzing the ontologies, closer to the knowledge domain of the MassConv system, we enriched semantics by extending the understandability of data by matching more general ontologies, the organization ontology (ORG-O) and the provenance ontology (PROV-O):

- provenance ontology (PROV-O) [33] describes the concepts related to the provenance in heterogeneous environments, and develops around three main classes: Agent, Entity, and Activity.
- organization ontology (ORG-O) [34] develops around the core class `org:Organization` which represents “a collection of people organized together into a community or other social, commercial or political structure”.

6.2. "Tacit" Knowledge Capture and Codification

MassConv "tacit" knowledge was codified in ontological entities, as classes and properties.

In order to capture "tacit" knowledge we have adopted the following method: (a) we have collected the most common parameters of software functions and we have selected those more relevant for the DigLib development; (b) we have matched, selected parameters with available written definitions, in the text documentation; (c) we created identifiers and written definitions for selected parameters; (d) we created the local ontology about the SDL system as the knowledge artifact for turning "tacit > encapsulated" into "codified" knowledge; (e) in the end we have matched local ontology to existing ontologies as already "codified" knowledge.

The conceptual elements, obtained by the method, have been codified as ontological entities, composing the ontology, about SDL, named On-SDL. A list of defined concepts in English and in Italian is at the URL https://sbs.uniroma1.it/test/data/vocabulary/itrousr-onsdl. The main classes representing the SDL "tacit" knowledge are:

- organizational collection (OrgColl)
  identifies the university organization, responsible for the selection and production of digital materials, and collects related descriptive data.
- digital collection (DigColl)
  is a DigRes which represents a specific set of DigRess. It collects the minimal data of belonging DigRes, the self-descriptive data for being identified and retrieved by the information system, and the data about the workflow.
- digital resource (DigRes)
  coherent and minimal descriptive information for an intellectual entity which is uniquely identified in the local management system.
- digital metadata object (DMO)
  data file in text format, firstly encoded in XML (encoding=UTF-8) and using metadata semantics, based on the metadata standards, adopted by the local system.
- digital content object (DCO)
  file of whatever digital format, representing an intellectual content or part of it.

More extensive and technical explanation about defined classes are available in the REMS 2018 paper [1].

In order to have an initial formalization about detected "tacit" knowledge concepts, we represented main concepts, roles and individuals in $\mathcal{ALC}$ the basic description logics [35].

We considered the $\mathcal{ALC}$ representation as the logic basement for detected ontological entities. The codification using description logics, supported us in selecting most relevant parameters and in representing relationships between them, thus to enrich "tacit" knowledge representation. Indeed by means of description logic representation we could connect the concept of ORG, which is responsible for data management. Figure 4 depicts the TBox $\mathcal{T}$ modeling the intentional knowledge [36], managed by the MassConv software. We can recognize the classes previously defined, and the logic model of the additional class UniversityORG which allows to codify connection to the ORGs responsible for data management and to establish matching with other ontologies.

The ontological entities of On-SDL TBox $\mathcal{T}$ are formatted in serif font and described as follows:

- $\mathsf{UniversityORG} \sqsubseteq \mathsf{UniversityDL}$: ORGs belonging to Sapienza can be represented in the SDL.
- $\mathsf{UniversityORG} \sqsubseteq \texttt{prov:Agent}$: Sapienza ORG is a type of PROV-O agent.
- $\texttt{org:Organization} \equiv \mathsf{UniversityORG}$: Sapienza ORG is a type of ORG.
- $\mathsf{DigitalObject} \sqsubseteq \mathsf{ContentObject} \sqcup \mathsf{MetadObject}$: DigObj is either a DCO or a DMO.

- ContentObject $\sqsubseteq$ DigitalObject: DCO is a type of DigObj.
- MetadObject $\sqsubseteq$ DigitalObject: DMO is a type of DigObj.
- ContentObject $\sqcup$ MetadObject $\sqsubseteq \bot$: nothing can be both DCO and DMO.
- `mods:ModsResource` $\sqsubseteq$ MetadObject: MODS resource is a type of DMO.
- `premis:Object` $\equiv$ DigitalObject: PREMIS object is equivalent to DigObj.
- DigitalCollection $\sqsubseteq$ DigitalResource: DigColl is a type of DigRes.

The $\mathcal{ALC}$ roles are expressed in domains and ranges, by using the existential quantifier $\exists$ and the universal quantifier $\forall$:

- UniversityORG manages DigitalResource: ORG manages at least one individual and all those individuals are Digress.
- UniversityDL aggregates DigitalResource: DigLib aggregates at least one individual and all those individuals are DigRess.
- DigitalCollection collects DigitalResource: DigColl collects at least one individual and all those individuals are DigRess.
- DigitalResource contains DigitalObject: DigRes contains at least one individual and all those individuals are DigObjs.

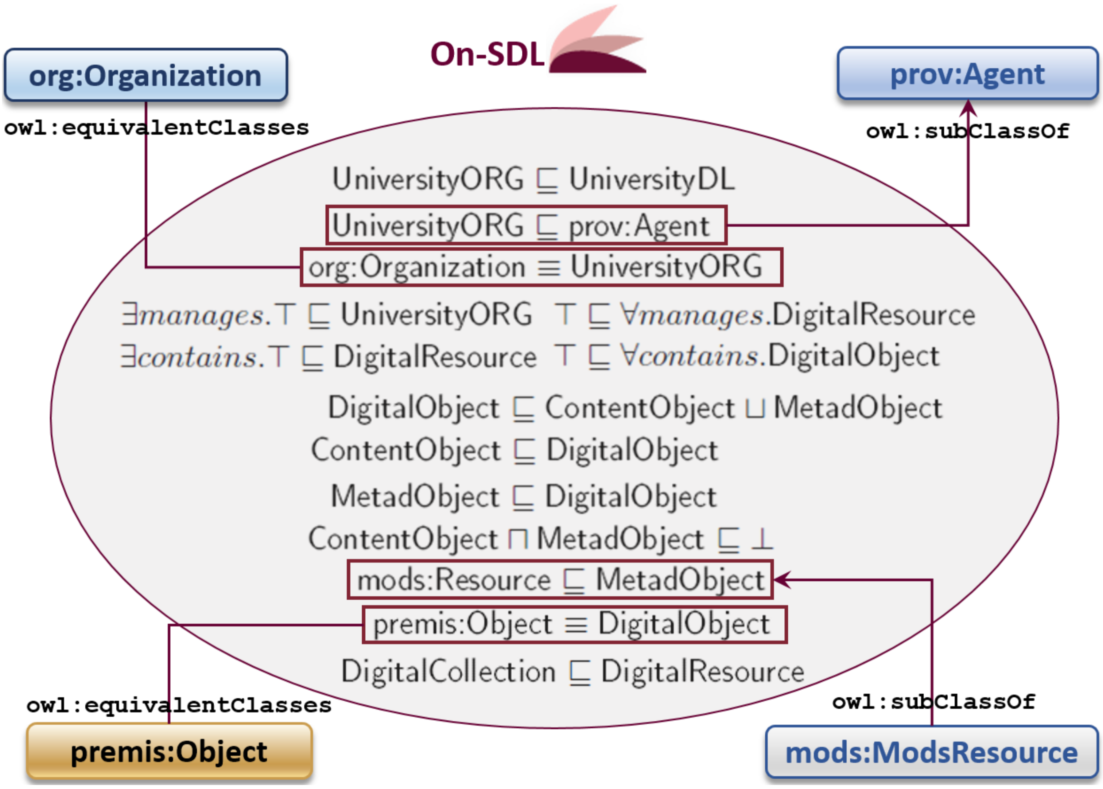
On-SDL
org:Organization
owl:equivalentClasses
prov:Agent
owl:subClassOf
UniversityORG $\sqsubseteq$ UniversityDL
UniversityORG $\sqsubseteq$ prov:Agent
org:Organization $\equiv$ UniversityORG
$\exists manages.\top \sqsubseteq$ UniversityORG $\top \sqsubseteq \forall manages.$DigitalResource
$\exists contains.\top \sqsubseteq$ DigitalResource $\top \sqsubseteq \forall contains.$DigitalObject
DigitalObject $\sqsubseteq$ ContentObject $\sqcup$ MetadObject
ContentObject $\sqsubseteq$ DigitalObject
MetadObject $\sqsubseteq$ DigitalObject
ContentObject $\sqcap$ MetadObject $\sqsubseteq \bot$
mods:Resource $\sqsubseteq$ MetadObject
premis:Object $\equiv$ DigitalObject
DigitalCollection $\sqsubseteq$ DigitalResource
owl:equivalentClasses
owl:subClassOf
premis:Object
mods:ModsResource

Figure 4. TBox of the local system representing the local tacit knowledge, and its matching with existing ontologies.

6.3. "Encapsulated" Knowledge Capture and Enrichment

Figure 1 shows the MassConv steps as yellow numbered arrows, witnessing how some DigLib knowledge is "encapsulated" into the system. We now generically describe the MassConv workflow steps, as follows:

1. organizational collection creation: Sapienza ORG is identified by an URI.
2. object acquisition: from a Sapienza ORG collects and stores multimedia objects and descriptive metadata, into a working area, assigns to new DigRess a URI, extending the ORG URI, associates related descriptive data (MODS), and computes or collects (if existing) preservation data (PREMIS).
3. mapping development: checks if there are new semantic entries, and in case, it learns the MassConv system with the new semantic, and a (manual) mapping development is requested.
4. object accessioning: from the *Acquisition* working area, copies multimedia objects in the SDL repository as DCOs, by propagating and extending DigRess' URIs to related DCOs.
5. collecting preservation metadata: collects and computes metadata about DCOs, necessary to the preservation of DCOs.
6. digital resource production: where required by the SDL ORG, DMOs and related DCOs are produced, according to the SDL XML metadata schemas $\mathcal{G}$, and conforming with DigRes content model.

We can observe that workflow steps, "encapsulated" in the software, as functions specifically performing that steps, adds more functional knowledge about the system and consequently, about data that are stored in the RDB.

The "encapsulated" knowledge was codified into a local ontology for describing the MassConv workflow (MCW-O). A representation of MCW-O and its connection to On-SDL, and its matching with existing ontologies, is showed in the Figure 5.

It is worth noting that the MCW-O classes, have been later modeled as subclasses of the `premis:Event`, an ontological entity of the PREMIS-OWL. Considering that the SDL adopted PREMIS standard and its controlled vocabularies (ID.LOC.GOV – Linked Data Service, Preservation vocabs, http://id.loc.gov/vocabulary/preservation.html), MCW-O classes are listed with other concepts taken from the PREMIS event vocabulary, and defined in English and in Italian (available at the URL https://sbs.uniroma1.it/test/data/vocabulary/itrousr-event).

7. System Knowledge Codified as a Linked Data Vocabulary

The LOV [8] are reusable vocabularies for the description of data on the Web. The LOV initiative gathers and makes visible indicators such as the interconnections between vocabularies and each vocabulary's version history, along with past and current editor (individual or organization).

In order to produce the LD provided with related LOVs, both local (On-SDL and MCW-O) and existing (PREMIS-OWL, MODS-O, PROV-O, ORG-O), we have followed the four LD principles defined in [37] and then further detailed in [5,38,39].

1. use URIs as names for things—"not just Web documents and digital content, but also real world objects and abstract concepts".
2. use HTTP URIs so that people can look up those names—"to identify objects and abstract concepts".
3. when someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)—"use of a single data model for publishing structured data on the web a simple graph-based data model that has been designed for use in the context of the web".
4. include links to other URIs, so that they can discover more things—"not only web documents, but any type of thing".

The data management turns to the semantic data management, by means of the application of the LD principles, where the management of URIs [40] is essential for the unambiguous reference to the SemWeb resources. Thus, the first step of detecting the tacit knowledge concepts into local ontologies, On-SDL and MCW-O, fulfills the first LD rule, because each ontological entity is identified by an URI.

Figure 5 shows a graph representation of the tacit knowledge detected by the method, described in the Section 6. On-SDL and MCW-O local ontologies are merged together and are represented as pink ellipses. Each ontological entity (classes and properties) belonging to the local ontology is identified by an URI prefixed by "onsdl:", the ontology expressing the knowledge about the DigLib resources. The MCW-O classes can be distinguished by the yellow tags, numbered according to the workflow steps, already described in the previous Section 6.3.

Existing ontologies are represented by differently colored ellipses, based on the ontology type. The matching assertions, declared by the On-SDL and expressing the founding knowledge about the DigLib system, drive the possibility of exposing LD that can be further interpreted by machines searching for predicates that belong to existing ontologies, PREMIS-OWL, MODS-O, PROV-O, ORG-O. Matching assertions represented in the Figure 5 witnesses also that the fourth LD principle is pursued and fulfilled.

The graphical representation was also published on the test website (still under development) at the URL https://sbs.uniroma1.it/test/sapienzadl/on-sdl.html, from where it is possible to access at the defined ontological entities.

It is worth noting that the MCW-O classes, as subclasses of `premis:Event`, link SKOS (SKOS primer [41], 4.2 Advanced Documentation Features https://www.w3.org/TR/skos-primer/#secadvanceddocumentation) vocabulary terms, already exhibited as LD, and providing definition in English and Italian.

The remaining second and third LD principles are fulfilled by producing the linked dataset and publishing it over the web protocol http (nevertheless, nowadays, the https protocol is mandatory). The test page to access at the prototype is at the URL https://sbs.uniroma1.it/test/data/vocabulary/.

The data will be published at the official page at the URL https://sbs.uniroma1.it/data/.

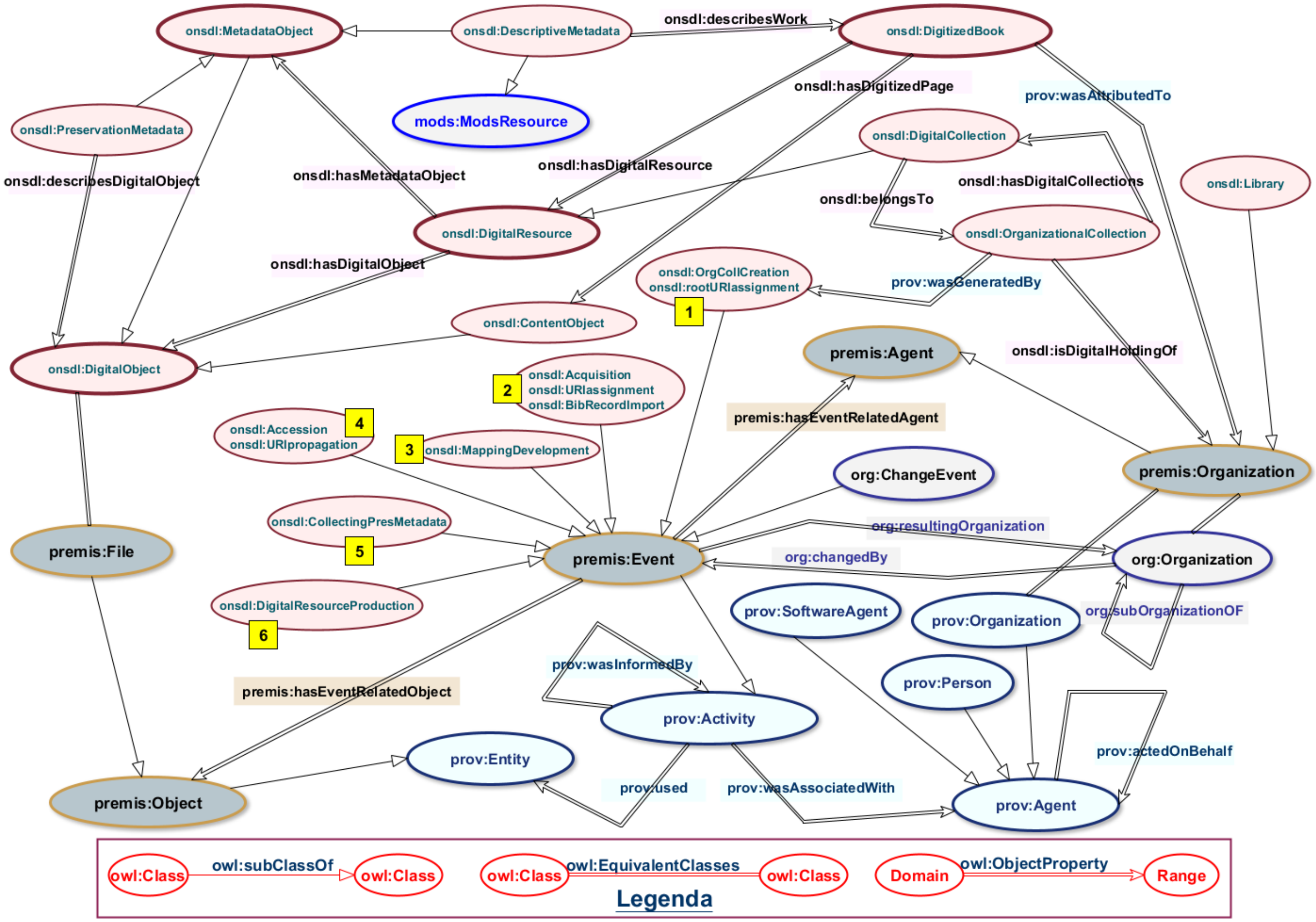

Figure 5. Overview of local and existing ontologies related to the ORG context, about data, of the DigLib case study.

8. Managing Data and Vocabularies, to Be Linked

Section 6 describes our approach for detecting the tacit knowledge underlying the MassConv system, as a DigLib system case study. In order to present our approach for generating the Linked Dataset using the ontological entities, defined by On-SDL and MCW-O ontologies, we now describe (1) the identification system, used for minting the URIs, (2) data at the source (the MassConv-RDB), (3) how we prepared the data mapping at the RDB source, toward system's ontological entities.

8.1. Semantic Data Management for URIs

As a semantic data management practice, we have reviewed and extended the identification method for identifying data stored into the MassConv-RDB, and for exhibiting it as LD. Taking into account the LD principles (see Section 7), we observe that the MassConv was natively equipped with the management of identifiers designed as URIs [42].m Figure 5 shows, indeed, specific events designed in the MassConv workflow that are focused on the management of URIs, specifically the steps 1—`rootURIassignment`, 2—`URIassignment` and 4—`URIpropagation`.

We review the URI management of the MassConv system, according to the main classes of matching ontologies.

- `premis:Agent` ⊑ `prov:Agent`

  The organizational perspective adopted for the data management of the MassConv has already identified all the Agents, participating to the production process of DigRess, and acted on behalf of the main Organization, the Sapienza University. According to the PREMIS-OWL ontology we collected the Agent data, distinguishing between `Organization`, `Person`, and `SoftwareAgent`, and identified them by adopting the same method for minting URIs.

  In order to make LD consumers aware about the source of data we used the global identifier, maintained by the Library of Congress (Library of Congress vocabularies, Cultural Heritage Organizations, http://id.loc.gov/vocabulary/organizations) for Cultural Heritage Organization, as the root identifier for minting local data URIs. Sapienza University is identified in the SemWeb space by the URL http://id.loc.gov/vocabulary/organizations/itrousr. In the Linked Dataset to be generated from the MassConv-RDB each resource URI is prefixed by that identifier (`itrousr-`), in order to allow a global recognizability of the exhibited resources. The following Turtle triples (Listing 1), expressed in Turtle syntax [43], provide an example of how Sapienza University as a cultural organization is represented as a LD resource, and how the `Library` is related to the Sapienza University:

  Listing 1: Turtle triples for a Sapienza's Library.

```
# Authoritative identifier for Sapienza University as a~
# Cultural Heritage Organization
:itrousr-RMS
a premis:Organization , premis:Agent ;
owl:sameAs <http://id.loc.gov/vocabulary/organizations/itrousr>.
#
# Library responsible for the book digitization
:itrousr-RMSAR-lib
a onsdl:Library premis:Organization , premis:Agent ;
prov:actedOnBehalfOf :itrousr-RMS .
```

- `premis:Object` ⊑ `prov:Entity`

  MassConv system, as a DigLib, produces DigRess' identifiers, based on the identifiers of the Italian National Bibliographic System (Anagrafe Biblioteche Italiane http://anagrafe.iccu.sbn.

it/opencms/opencms/), that reflect the organizational structure of the Sapienza University libraries. A DigRes managed by the MassConv, is uniquely identified by the root identifier of the Sapienza's library, as the holding organization of the physical books, that have been digitized and reproduced as DigRess. For example, the Sapienza Architecture Library is identified by "RMSAR", the root identifier, used by the MassConv system, for identifying each DigRes to be produced, and which belongs to the identified library. (The reader can notice that the phrase segments underlined in the previous paragraph matches with some classes and properties defined by the ontology.) In 2013, the MassConv system produced 1067 `MetadataObject` as XML files (1057 `DigitalResource` and 10 `DigitalCollection`), that were published as Open Data since 2017 at the URL, https://sbs.uniroma1.it/sapienzadl/. For example the `DigitalCollection` of the Sapienza Architecture Library is retrievable at the URL https://sbs.uniroma1.it/data/opendata/itrousr-od_2017-SDL_2013_RMSAR. The XML file (a `MetadataObject`) of the Sapienza Architecture Library collection, the `DigitalCollection` identified by "RMSAR", is at the URL https://sbs.uniroma1.it/openDataSets/sdl2013/METSXML/RMSAR/RMSAR.xml, while the `MetadataObject` of a belonging book (a `DigitalResource`) is retrievable at the URL https://sbs.uniroma1.it/openDataSets/sdl2013/METSXML/RMSAR/RMSAR_00000025.xml. According to the SemWeb vision, these objects, are informational (XML/HTML documents) resources for human consumption, while informational LD resources, are for machine [44] consumption, as well as the knowledge about "things" identified and codified in a SemWeb language (OWL).
When an URI is already available for identifying `OrganizationalCollection` `DigitalCollection`, `DigitalResource` or `DigitalObject`, etc., the identifier is only prefixed by `itrousr-`, for distinguishing the URI of DigRess (also used for historical reasons) and the URI used for LD.

Listing 2 shows prototype's Turtle triples, that are related to the main classes of the On-SDL ontology, which defines "things" mainly managed by the DigLib system. The following Turtle triples express (1) why descriptive data are managed by the system: data describes the intellectual content of a digitized book, managed by the DigLib system; (2) how data is managed, and structurally collected in the context of the holding organization. Thus the organizational context is identified, and expressed, for each system's object of concern.

Listing 2: Turtle triples, expressing the organizational context of a Sapienza Digital Resource: https://sbs.uniroma1.it/openDataSets/sdl2013/METSXML/RMSAR/RMSAR_00000025.xml.

```
#=====================> OrganizationalCollection class
:itrousr-RMSAR-org_coll
a onsdl:OrganizationalCollection ;
onsdl:hasDigitalCollection :itrousr-RMSAR ;
onsdl:hasDigitalCollection :itrousr-RMSAR_PRODIGI ;
onsdl:hasDigitalCollection :itrousr-RMSAR_SEVERATI ;
onsdl:isDigitalHoldingOf :itrousr-RMSAR-lib .
#=====================> DigitalCollection class
:itrousr-RMSAR_PRODIGI
a onsdl:DigitalCollection , onsdl:DigitalResource ;
belongsTo :itrousr-RMSAR-org_coll ;
:itrousr-RMSAR_SEVERATI
a onsdl:DigitalCollection , onsdl:DigitalResource ;
belongsTo :itrousr-RMSAR-org_coll ;
:itrousr-RMSAR
a onsdl:DigitalCollection , onsdl:DigitalResource ;
belongsTo :itrousr-RMSAR-org_coll ;
#=====================> DescriptiveMetadata class
```

```
:itrousr-DescriptiveMetadata_VEAE007681
a onsdl:DescriptiveMetadata , onsdl:MetadataObject , prov:Entity , mods:ModsResource ;
onsdl:describesWork :itrousr-RMSAR_RISFLORIANI ;
prov:wasGeneratedBy :itrousr-bibRecordImport_2019-00003504 ;
prov:wasAttributedTo :itrousr-sdl_ap_000070 .
#=====================> DigitizedBook class
:itrousr-RMSAR_RISFLORIANI
a onsdl:DigitizedBook , prov:Entity ;
onsdl:hasDigitalResource :itrousr-RMSAR_00000025 ;
onsdl:hasDigitizedPage :itrousr-RMSAR_00000025_0001-jpg ;
prov:wasAttributedTo :itrousr-sdl_ap_000028 .
#=====================> DigitalResource class
:itrousr-RMSAR_00000025
a onsdl:DigitalResource , prov:Entity ;
prov:wasGeneratedBy :itrousr-URIassignment_2019-00008227 ;
prov:wasAttributedTo :itrousr-sdl_as_000072 .
```

Listing 3 shows Turtle triples, related to the classes of the On-SDL ontology, defining "things" mainly managed by the DigLib system. The following Turtle triples provide details about how a type of entity, defined by an existing ontology (PREMIS_OWL) is further identified in the local system, in relation to its management. In the sample, `Accession` and `DigitalResourceProduction`.

Listing 3: Turtle triples expressing how the system terminologically distinguishes `premis:Object`.

```
#=====================> ContentObject --> DigitalObject class
:RMSAR_00000025_0007-tif
a onsdl:ContentObject , onsdl:DigitalObject , premis:Object ;
prov:wasGeneratedBy :itrousr-Accession_2019-00003504 ;
prov:wasAttributedTo :itrousr-sdl_as_000072 .
#
# [amount of ContentObject is 341 jpg and 342 tiff]
#=====================> MetadataObject --> DigitalObject class
#
:RMSAR_00000025-xml
a onsdl:MetadataObject , onsdl:DigitalObject , premis:Object ;
prov:wasGeneratedBy :itrousr-DigitalResourceProduction_2019-00003504 ;
prov:wasAttributedTo :itrousr-sdl_as_000072 .
#=====================> PreservationMetadata ==> DigitalObject class
:RMSAR_00000025-preserve
a onsdl:PreservationMetadata , onsdl:MetadataObject , premis:Object ;
prov:wasGeneratedBy :itrousr-DigitalResourceProduction_2019-00003504 ;
prov:wasAttributedTo :itrousr-sdl_as_000072 .
```

- `premis:Event` ⊑ `prov:Activity`

  Listing 4 shows the most representative Turtle triples, in the prototype, for expressing how a DigRes has been obtained.

  It is worth noticing that this is the last step performed by the MassConv system, all previous steps are identified and related to this event by means of `prov:wasInformedBy`. The digitized pages (from one to 341 for JPEG format, and from one to 342 for TIFF format), and the incoherence between the number of TIFF files and the JPEG, shows that some event was not completed, determining the object missing.

Listing 4: Turtle triples expressing how the Sapienza Digital Resource was produced: https://sbs.uniroma1.it/openDataSets/sdl2013/METSXML/RMSAR/RMSAR_00000025.xml.

```
#================> DigitalResourceProduction event
:itrousr-DigitalResourceProduction_2019-00003504
```

```
a premis:event, prov:Activity ;
prov:used :itrousr-DescriptiveMetadata_VEAE007681
prov:used :RMSAR_00000025_0001-jpg ;
[...]
prov:used :RMSAR_00000025_0341-jpg ;
prov:used :RMSAR_00000025_0001-tif ;
[...]
prov:used :RMSAR_00000025_0342-tif ;
prov:used :RMSAR_00000025-xml ;
prov:used :RMSAR_00000025-preserve
prov:wasInformedBy :itrousr-OrgCollCreation_2019-00003504 ;
prov:wasInformedBy :itrousr-RootURIassignment_2019-00003504 ;
prov:wasInformedBy :itrousr-BibRecordImport_2019-00003504 ;
prov:wasInformedBy :itrousr-Acquisition_2019-00003504 ;
prov:wasInformedBy :itrousr-URIassignment_2019-00008227
prov:wasInformedBy :itrousr-MappingDevelopment_2013-00000568 ;
prov:wasInformedBy :itrousr-Accession_2019-00003504 ;
prov:wasInformedBy :itrousr-URIpropagation_2019-00003504 ;
prov:wasInformedBy :itrousr-CollectingPresMetadata_2019-00003504 ;
prov:wasAssociatedWith :itrousr-sdl_as_000072 ;
prov:startedAtTime "2013-10-11T20:36:47Z"^^xsd:dateTime ;
prov:endedAtTime   "2013-12-11T21:14:55Z"^^xsd:dateTime .
```

8.2. Semantic Data Management and the MassConv Data Source

The data management of the MassConv DigLib system is based on the MySQL (MySQL, https://www.mysql.com/) an open source relational database management system (RDBMS).

Table 1 shows the list of the MassConv-RDB tables and the corresponding amount of Rows.

It is worth noticing that the table bridge_identifiers contains the identifiers uniquely assigned to each descriptive bibliographic record (converted in `DescriptiveMetadata`) provided by the holding `Library`, having the main responsibility for the existence of DigRess and related DigColls. The management of identifiers was indeed one of the most important requirements for the data management, performed by the MassConv system. The "[OrgColl]" prefix represents the set of identifiers assigned to each `OrganizationalCollection`, because the MassConv-RDB tables are horizontally partitioned, according to each `OrganizationalCollection` known by the system. In the data sample, the prefix "[OrgColl]" represents the set of identifiers assigned to 46 Sapienza Libraries, that produced digitized materials.

The RDB table [OrgColl]_file_objects contains data for 27,808 digitized books, containing 36,358,076* DigObjs. The star sign reminds the reader that the data is approximate, because the amount of digitized books has increased to 55,372. The correct amount of DigObjs is under computation. The histogram of Figure 6 shows the current status of `OrganizationalCollection` with the distribution of 48,112 `DigitalResources`, that have completed the second step of the MassConv workflow (see Section 6.3). The histogram of Figure 7 shows the `DigitalCollection` distribution of 1494 `DigitalResources`, that are collected, by URL reference, thus, they are a subset of the 48,112 `DigitalResources`, and it can be inferred that only those 1494 `DigitalResource`s have been processed by all steps of the MassConv workflow (see Section 6.3). The histogram of Figure 8 shows the `OrganizationalCollection` distribution of 55,372 `DigitalResource`s produced by the GoogleBooks digitization project, undertaken by Sapienza from 2012, until 2017. The majority of 48,112 `DigitalResource`s were processed from GoogleBooks, and more than 1000 are still under identification process. The histogram of Figure 9 shows the `OrganizationalCollection` distribution of 36,358,076 `ContentObjects` produced by the GoogleBooks digitization project until the 2015.

The actual number of the `ContentObject` produced by the project is a work in progress. The list of `DigitalResource`s 48,112 and 1494, distributed over 57 `DigitalCollection`s is reachable at the URL https://sbs.uniroma1.it/test/sapienzadl/it/itrousr-DigColl_list. The list is the access URL for the informational resource, that provide a human and machine access to the corresponding linked dataset.

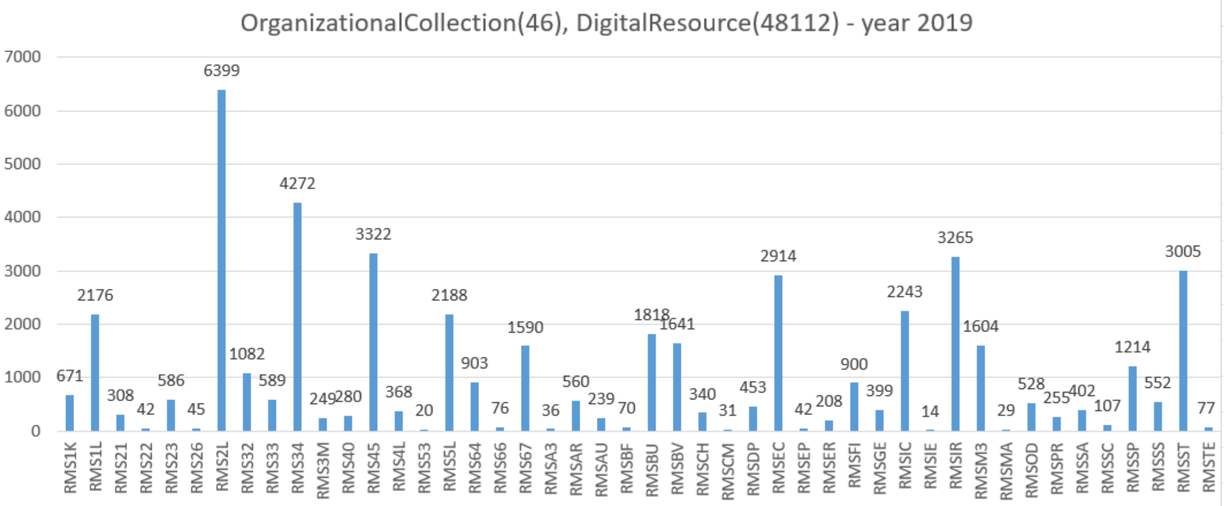

Figure 6. Collection of 48,112 digital resources—46 organizational collection—year 2019.

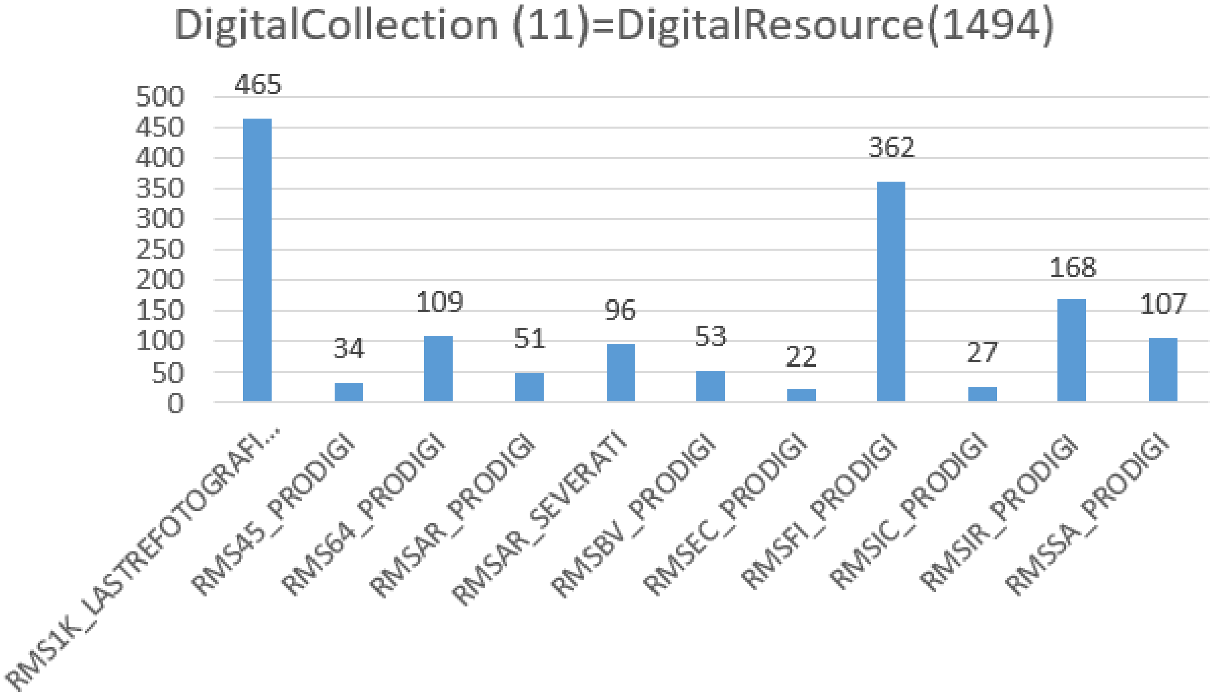

Figure 7. Collection of 1494 digital resources—11 digital collection—year 2019.

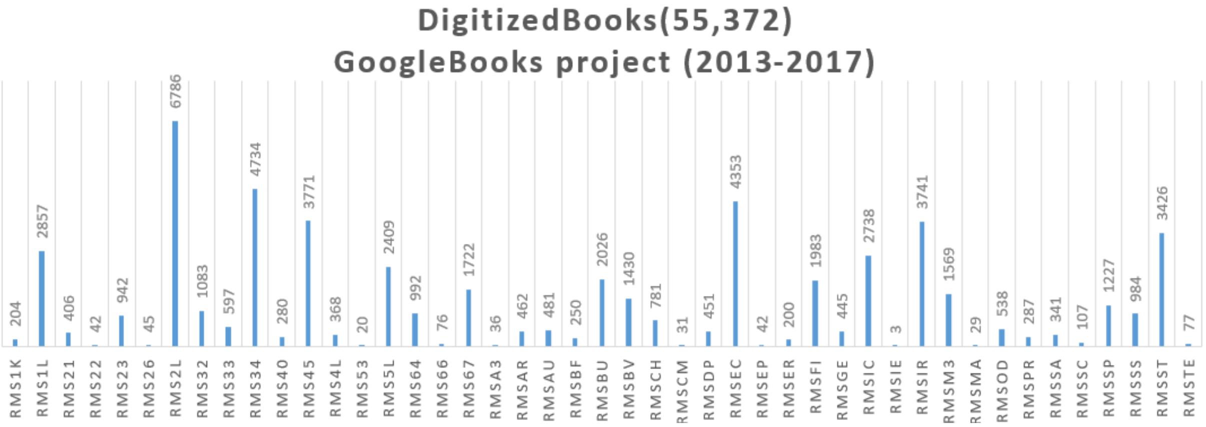

Figure 8. Collection of 55,372 digitized books by Google.

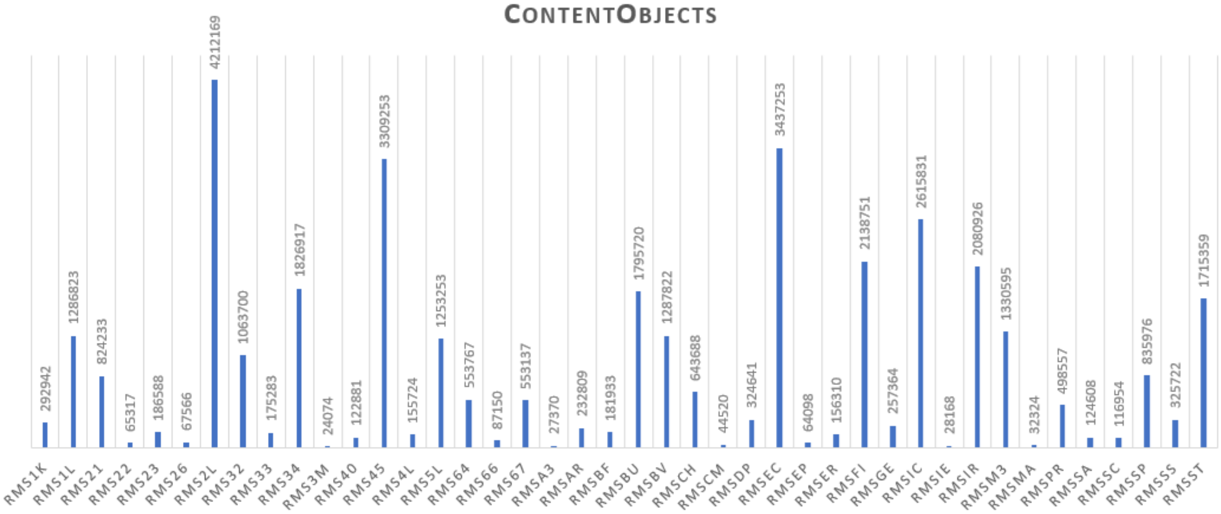

Figure 9. Collection of 36,358,076 content objects per 42 collection—year 2015.

8.3. Semantic Data Management and the Data Source Mapping

Figure 10 gives a simplified view of the MassConv-RDB schema. As a semantic data management practice, we have stored a table for the local ontology data (S_0), and related terms have been mapped toward the relevant RDB tables, identified as the data sources for generating LD. The data source S_{10} is computed by collecting processed time, stored in the tables, prefixed by [OrgColl] (see Table 1). If a data exists in a table, associated with an event, the data is computed and generated as an LD resource. For example, if we find the URI of a `DigitalResource` (i.e., https://sbs.uniroma1.it/openDataSets/sdl2013/METSXML/RMSAR/RMSAR_00000025.xml) in the table S_6 it means that the `DigitalResourceProduction` has happened for that URI, thus the data of preceding events in the workflow can be also computed. As a semantic data management practice, a simplified mapping between the data source and the ontology classes, drives also the derivation RDB tables' relationships. The RDB relationships indeed, are not depicted, because the properties, defined by the local, and matching ontologies, will drive the LD generation, and unveil the RDB relationships, in a semantic way.

Table 1. The MassConv-RDB tables (* the data is only a sample of DigitalObject collected in June 2015 for 27,808 DigitizedBooks, the current number of DigitalObject, is under computation).

Table Name	Description	Rows
org_coll	University organizational collection	157
collections	Digital collections managed by the system	57
agent_roles	Information about the roles of agents in the events	283
agents	Agents involved in the DigLib system	99
acquisition	Workflow event for the identification of digitized materials	1026+46,125
[OrgColl]_resources	Digital resource index	96+48,112
[OrgColl]_file_objects	Content objects' inventory	191+193,021
[OrgColl]_file_objects	Content objects' inventory (2015)	36,358,076*
[OrgColl]_metadata_objects	Metadata objects' inventory	191+193,021
[OrgColl]_materials	Digitized objects as submitted by provider	193+224,503+ 36,358,076*
[OrgColl]_Events	Workflow events	--

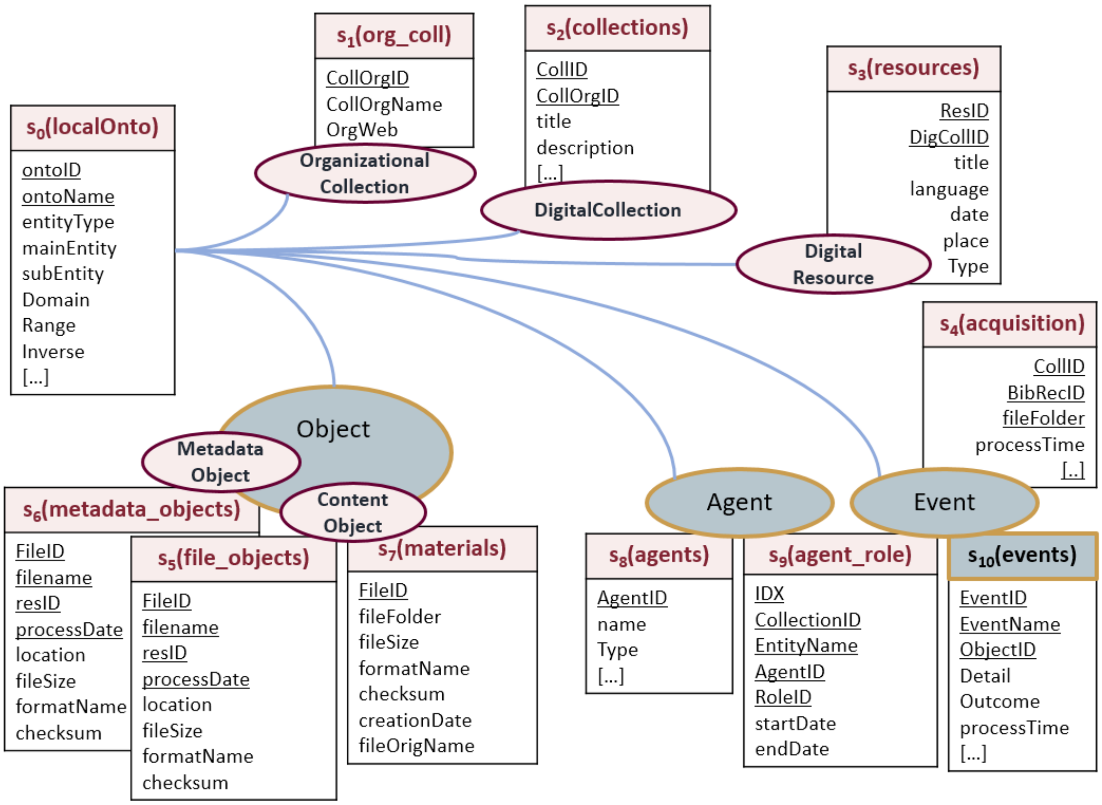

Figure 10. MassConv relational database (RDB) schema without relationships.

9. Implementing on the Data Management Case Study and First Results

Table 2 show the amount of named individuals [45,46] (the instances of "things" defined for describing the system, and identified by URIs), that we have detected as belonging to a class and identified with URIs. The On-SDL classes are ordered following the workflow management logic of the system: named individuals of Libraries involved in the digitization projects are 46 of 65 libraries of Sapienza, to which 46 correspond to OrganizationalCollection of 157 possible organizational collections of Sapienza (Sapienza is composed by 157 organizations, departments, libraries, museums, etc.). From the 46 OrganizationalCollections, 57 DigitalCollection have been created by gathering URI references about specific sets of DigitalResources based on a scientific selection (i.e., ancient books, phd thesis, etc.).

A collection of 48,265 DescriptiveMetadata, corresponding to 48,265 ModsResource, describe only 48,265 out of 55,372 DigitizedBooks by Google and 1805 by libraries and 57 DigitalCollections.

It is worth noting that the first 1067 (**961** + **96** + **10**) DigRess produced, were published in the SDL web since 2013, and as open data. Thus the boldface distinguishes these DigRess, because submitted to all workflow events, defined by the MCW-O classes, the **193,021** number highlights related ContentObjects, and PreservationMetadata involved in the process.

The 48,265*2 MetadataObject comprehends 48,265 DescriptiveMetadata and 48,265 PreservationMetadata. Resources added at a later time, progressively increased the amount of resources at a specified step of the workflow. As such, it is possible to infer the status of resources in the workflow, managed by the system, at certain point in time.

Table 3 shows the corresponding amount of resources for each property defined in the local ontology.

Table 2. Named individuals identified in the massive conversion (MassConv)-relational database (RDB). Bolded numbers represent DigRes data produced since 2013, other collected data shows the evolution and the increment of managed DigRess, that still have to be produced.

Onto	OntoName	Named Individuals [45,46]	SuperClass
onsdl	Library	46/65	
onsdl	OrganizationalCollection	46/157	
onsdl	DigitalCollection	46 + 11	
onsdl	DescriptiveMetadata	**961** + **96** + **10** + 46,125 + 57	MetadataObject
mods	ModsResource	48,265	
onsdl	DigitizedBook	1805 + 55,372	Book
onsdl	DigitalObject	(48,265*2) + **193,021** + 31,296 + 36,358,076*	
onsdl	ContentObject	**193,021** + 31,296 + 36,358,076*	DigitalObject
onsdl	PreservationMetadata	**193,021** + 31,296 + 36,358,076*	MetadataObject
onsdl	MetadataObject	48,265*2	DigitalObject
onsdl	DigitalResource	**961** + **96** + **10** + 1026 + 46,125 + 47	
mcwo	OrgCollCreation	157	Event
mcwo	RootURIassignment	46	Event
mcwo	Acquisition	1,026 + 55,372	Event
mcwo	URIassignment	57 + 48,265	Event
mcwo	BibRecordImport	2,214,190	Event
mcwo	MappingDevelopment	143	Event
mcwo	Accession	**961** + **96** + **10**	Event
mcwo	URIpropagation	**193,021** + 31,296 + 36,358,076*	Event
mcwo	CollectingPresMetadata	**193,021** + 31,296 + 36,358,076*	Event
mcwo	DigitalResourceProduction	**961** + **96** + **10**	Event

Table 3. Named individuals identified in the MassConv-RDB.

Onto	OntoName	Named Individuals [45,46]	Domain ⇒ Range
onsdl	`hasDigitalCollection`	46 ⇒ 57	`OrganizationalCollection` ⇒ `DigitalCollection`
onsdl	`hasDigitalObject`	48,265 ⇒ (48,265*2) + **193,021** + 31,296 + 36,358,076*	`DigitalResource` ⇒ `DigitalObject`
onsdl	`hasDigitalResource`	57,177 ⇒ 48,265	`DigitizedBook` ⇒ `DigitalResource`
onsdl	`hasDigitizedPage`	57,177 ⇒**193,021** + 31,296+36,358,076*	`DigitizedBook` ⇒ `ContentObject`
onsdl	`hasMetadataObject`	48,265 ⇒ 48,265*2	`DigitalResource` ⇒ `MetadataObject`
onsdl	`isDigitalHoldingOf`	46 ⇒ 46	`OrganizationalCollection` ⇒ `Library`
onsdl	`belongsTo`	57 ⇒ 46	`DigitalCollection` ⇒ `OrganizationalCollection`
onsdl	`describesDigitalObject`	**193,021** + 31,296 + 36,358,076* ⇒ **193,021** + 31,296 + 36,358,076*	`PreservationMetadata` ⇒ `DigitalObject`
onsdl	`describesWork`	**961** + **96** + **10** + 46,125 + 57 ⇒ **961** + **96**+**10** + 46,125 + 57	`DescriptiveMetadata` ⇒ `DigitizedBook`

10. Limitations, Conclusions and Future Developments

In this paper, we have presented an experimental method for capturing, and expressing "tacit" knowledge, in the Linked Dataset, prototyped from a DigLib system case study. Due to its experimental nature, the method should be evaluated in other InfSys, and the quality of resulting LD should be assessed. The method points to turn data management practices, into "semantic data management practices", where the system knowledge is managed as a data, along with other managed data. The adoption of such practices improves the way SemWebTech is implemented in a legacy InfSys, by developing a mindset of knowledge workers, oriented toward the management of LD pillar elements, the URI and LOV, and by enriching Linked Datasets, produced and published, with the tacit knowledge of data organizational context. LD semantics thus convey the knowledge necessary to understand "why data was created" and "how data was managed", and to re-use or re-manage data in a proper way.

The interpretation of data is facilitated both for humans, that might be unloaded by long and discontinue searches of additional information, and for machines that could re-use data with more accuracy. Semantic data management practices, even being mostly manual when started to be applied, can use automatic tools for extracting "encapsulated" knowledge, from software and data management platforms. Instead, capturing "tacit" knowledge, not "codified", nor "encapsulated", requires a mindset of the system developers and data managers, oriented to convey the data meaning by means of SemWebTech, consequently, the method is proposed as a training for developing such a mindset.

Achieving the goal of increasing a better understanding, and a proper reuse of LD is strongly hindered by the quality of data, and "capturing the organizational knowledge about the meaning of the data" is not a straightforward task and still mostly manual. Indeed, it is well-known in literature the problem of ontology engineering process, which is defined as labor-intensive, error-prone and time-consuming.

The "tacit" portion of an InfSys knowledge plays a key role in this process, that influence performance of data stakeholders, thus it is important to remind that knowledge about data and its conceptualization and formalization in a SemWeb language, should be considered as a part of the software development task. Establishing a software development mindset, that consider the knowledge resources as part of the work, would improve software construction and maintenance [16], and make system more sustainable and more flexible to the organizational changes.

The more InfSys data increases in amount and complexity, and the more increases, the need of using machines for managing InfSys based on computable and interpretable data.

Proposed semantic data management increases the curation of data in the InfSys border, influencing the management of data with the global view of the "open world assumption", that applies when a system has incomplete information. It is well known that the ideal RDB is well-designed, and its data consistency and coherence is maintained for its whole life-cycle, but the reality is different. Changes in the functional requirements, software evolution and organizational changes have always impact on RDB data and schema, causing incomplete information.

The RDB incomplete information can be inferred from the available data, and related knowledge data. The application of LD principles, in the data management practices, allows data managers, to deal also with the inconsistency of data.

As future developments, we will publish a local ontology in the LOV registry (linked open vocabulary, https://lov.linkeddata.es/dataset/lov/), we will refine the semantic data management method, and we will test the workflow for capturing tacit knowledge.

As an immediate development, we are going to publish the Linked Datasets, related to the published digital resource, and progressively we will publish remaining resource that the MassConv will manage. The test website https://sbs.uniroma1.it/test/sapienzadl/it/itrousr-DigColl_list, is the place where Linked Dataset prototyped is published, once the test is completed, the Linked Dataset will be published at the URL https://sbs.uniroma1.it/data/.

The implementation of a SPARQL end point is also considered a further development, as well as to assess the quality of the Linked Dataset according to the metrics defined by Radulovic et al. [24].

Author Contributions: Conceptualization, A.D.I.; Investigation, A.D.I.; Methodology, A.D.I. and M.S.; Writing—original draft, A.D.I.; Writing—review and editing, A.D.I. and M.S.

Funding: This research received no external funding.

Acknowledgments: We thank the Sapienza Library System for providing bibliographic data.

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

References

1. Di Iorio, A.; Schaerf, M. Addressing the tacit knowledge of a digital library system. In Proceedings of the REMS 2018, Multidisciplinary Symposium on Computer Science and ICT, Stavropol, Russia, 15 October 2018; pp. 41–51.
2. Zaveri, A.; Rula, A.; Maurino, A.; Pietrobon, R.; Lehmann, J.; Auer, S. Quality assessment for linked data: A survey. *Semant. Web* **2015**, *7*, 63–93. [CrossRef]
3. Smith-Yoshimura, K. Analysis of International Linked Data Survey for Implementers. *D-Lib Mag.* **2016**, *22*. [CrossRef]
4. Tosaka, Y.; Park, J.R. Continuing Education in New Standards and Technologies for the Organization of Data and Information. *Libr. Res. Tech. Serv.* **2018**, *62*, 4–15.
5. Hyland, B.; Atemezing, G.; Villazón-Terrazas, B. Best Practices for Publishing Linked Data. Available online: https://www.w3.org/TR/ld-bp/ (accessed on 7 May 2019).
6. Hyland, B.; Atemezing, G.; Pendleton, M.; Srivastava, B. Linked Data Glossary. Available online: http://www.w3.org/TR/ld-glossary/ (accessed on 7 May 2019).
7. Cyganiak, R.; Wood, D.; Lanthaler, M.; Klyne, G.; Carroll, J.J.; McBride, B. RDF 1.1 concepts and abstract syntax. *W3C Recomm.* **2014**, *25*. Available online: https://www.w3.org/TR/rdf11-concepts/ (accessed on 7 May 2019).
8. Vandenbussche, P.Y.; Atemezing, G.A.; Poveda-Villalón, M.; Vatant, B. Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. *Semant. Web* **2017**, *8*, 437–452. [CrossRef]

9. Polanyi, M. *The Tacit Dimension*; University of Chicago Press: Chicago, IL, USA, 2009.
10. Hayes, P.J.; Patel-Schneider, P.F. RDF 1.1 Semantics. *W3C Recomm.* **2014**, *25*, 7–13.
11. Sahoo, S.S.; Halb, W.; Hellmann, S.; Idehen, K.; Thibodeau, T., Jr.; Auer, S.; Sequeda, J.; Ezzat, A. A survey of current approaches for mapping of relational databases to RDF. *W3C RDB2RDF Incubator Gr. Rep.* **2009**, *1*, 113–130.
12. Bizer, C.; Cyganiak, R. D2r server-publishing relational databases on the semantic web. In Proceedings of the 5th International Semantic Web Conference, Athens, GA, USA, 5–9 November 2006; Volume 175.
13. Beneventano, D.; Bergamaschi, S.; Sorrentino, S.; Vincini, M.; Benedetti, F. Semantic annotation of the CEREALAB database by the AGROVOC linked dataset. *Ecol. Inf.* **2015**, *26*, 119–126. [CrossRef]
14. Alavi, M.; Leidner, D.E. Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Q.* **2001**, *25*, 107–136. [CrossRef]
15. Ward, J.; Aurum, A. Knowledge management in software engineering-describing the process. In Proceedings of the Software Engineering Conference, Melbourne, Australia, 13–16 April 2004; pp. 137–146.
16. De Vasconcelos, J.B.; Kimble, C.; Carreteiro, P.; Rocha, Á. The application of knowledge management to software evolution. *Int. J. Inf. Manag.* **2017**, *37*, 1499–1506. [CrossRef]
17. Rowley, J.E. The wisdom hierarchy: Representations of the DIKW hierarchy. *J. Inf. Sci.* **2007**, *33*, 163–180. [CrossRef]
18. Wiig, K.M. *Knowledge Management Foundations: Thinking about Thinking-How People and Organizations Represent, Create, and Use Knowledge*; Schema Press: Arlington, VA, USA, 1994.
19. Evans, M.; Dalkir, K.; Bidian, C. A holistic view of the knowledge life cycle: the knowledge management cycle (KMC) model. *Electron. J. Knowl. Manag.* **2015**, *12*, 47.
20. Boisot, M.H. *Knowledge Assets: Securing Competitive Advantage in the Information Economy*; OUP Oxford: Oxford, UK, 1998.
21. Grant, R.M. Toward a knowledge-based theory of the firm. *Strateg. Manag. J.* **1996**, *17*, 109–122. [CrossRef]
22. Choo, C.W. The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. *Int. J. Inf. Manag.* **1996**, *16*, 329–340. [CrossRef]
23. Van den Berg, H.A. Three shapes of organisational knowledge. *J. Knowl. Manag.* **2013**, *17*, 159–174. [CrossRef]
24. Radulovic, F.; Mihindukulasooriya, N.; García-Castro, R.; Gómez-Pérez, A. A comprehensive quality model for linked data. *Semant. Web* **2018**, *9*, 3–24. [CrossRef]
25. Catarci, T.; Di Iorio, A.; Schaerf, M. The Sapienza Digital Library from the Holistic Vision to the Actual Implementation. *Procedia Comput. Sci.* **2014**, *38*, 4–11, doi:10.1016/j.procs.2014.10.002. [CrossRef]
26. Lenzerini, M. Data Integration: A Theoretical Perspective. In Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Madison, WI, USA, 2–6 June 2002; ACM: New York, NY, USA, 2002; pp. 233–246. doi:10.1145/543613.543644. [CrossRef]
27. PREMIS Editorial Committee. PREMIS Data Dictionary for Preservation Metadata, Version 3.0. 2015. Available online: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf (accessed on 7 May 2019).
28. Consultative Committee for Space Data. *Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book*; CCSDS Press: Washington, DC, USA, 2012.
29. Denenberg, R.; Guenther, R.; Han, M.J.; Luna Lucero, B.; Mixter, J.; Nurnberger, A.L.; Pope, K.; Wacker, M. *Making MODS to Linked Open Data: A Collaborative Effort for Developing MODS/RDF*; Columbia University Press: New York, NY, USA, 2014. doi:10.7916/D8125QTC.
30. Coppens, S.; Verborgh, R.; Peyrard, S.; Ford, K.; Creighton, T.; Guenther, R.; Mannens, E.; Van de Walle, R. Premis owl. *Int. J. Digit. Libr.* **2015**, *15*, 87–101. [CrossRef]
31. Di Iorio, A.; Caron, B. PREMIS 3.0 Ontology: Improving Semantic Interoperability of Preservation Metadata. In Proceedings of the 13th International Conference on Digital Preservation, Bern, Switzerland, 3–6 October 2016; pp. 32–36.
32. Blair, C.; Bountouri, L.; Caron, B.; Cowles, E.; Di Iorio, A.; Guenther, R.; McLellan, E.; Roke, E.R. 305.2 PREMIS 3 OWL Ontology: Engaging sets of linked data—Award Winner: Best Short Paper. In Proceedings of the 15th International Conference on Digital Preservation, Boston, MA, USA, 24–27 September 2018. doi:10.17605/OSF.IO/E8VJ6.

33. Lebo, T.; Sahoo, S.; McGuinness, D.; Belhajjame, K.; Cheney, J.; Corsar, D.; Garijo, D.; Soiland-Reyes, S.; Zednik, S.; Zhao, J. *PROV-O: The Prov Ontology: W3C Recommendation, 30 April 2013*; World Wide Web Consortium: Cambridge, MA, USA, 2013.
34. Reynolds, D. The Organization Ontology. 2014. Available online: http://www.w3.org/TR/vocab-org (accessed on 7 May 2019).
35. Nardi, D.; Brachman, R.J. An introduction to description logics. In *Description Logic Handbook*; Cambridge University Press: Cambridge, UK, 2003; pp. 1–40.
36. Baader, F. *The Description Logic Handbook: Theory, Implementation and Applications*; Cambridge University Press: Cambridge, UK, 2003.
37. Berners-Lee, T. Linked Data-Design Issues. 2006. Available online: http://www.w3.org/DesignIssues/LinkedData.html (accessed on 7 May 2019).
38. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data-the story so far. In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*; IGI Global: Hershey, PA, USA, 2009; pp. 205–227.
39. Heath, T.; Bizer, C. Linked data: Evolving the web into a global data space. *Synth. Lect. Semant. Web Theory Technol.* **2011**, *1*, 1–136. [CrossRef]
40. Berners-Lee, T. Cool URIs don't change. 1998. Available online: https://www.w3.org/Provider/Style/URI (accessed on 7 May 2019).
41. Miles, A.; Bechhofer, S. SKOS simple knowledge organization system reference. *W3C Recomm.* **2009**, *18*, W3C.
42. Di Iorio, A.; Schaerf, M. Identification Semantics for an Organization, establishing a Digital Library System. In Proceedings of the 4th International Workshop on Semantic Digital Archives (SDA 2014), Oxford, UK, 12 May 2014.
43. Beckett, D.; Berners-Lee, T.; Prud'hommeaux, E.; Carothers, G. *RDF 1.1 Turtle*; World Wide Web Consortium: Cambridge, MA, USA, 2014.
44. Michel, F.; Montagnat, J.; Zucker, C.F. A Survey of RDB to RDF Translation Approaches and Tools. Available online: https://hal.archives-ouvertes.fr/hal-00903568/file/Rapport_Rech_I3S_v2_-_Michel_et_al_2013_-_A_survey_of_RDB_to_RDF_translation_approaches_and_tools.pdf (accessed on 7 May 2019).
45. Motik, B.; Patel-Schneider, P.F.; Parsia, B.; Bock, C.; Fokoue, A.; Haase, P.; Hoekstra, R.; Horrocks, I.; Ruttenberg, A.; Sattler, U.; et al. OWL 2 web ontology language: Structural specification and functional-style syntax. *W3C Recomm.* **2009**, *27*, 159.
46. Schneider, M.; Carroll, J.; Herman, I.; Patel-Schneider, P.F. OWL 2 Web Ontology Language RDF-Based Semantics. 2009. Available online: https://www.w3.org/TR/owl2-rdf-based-semantics/ (accessed on 7 May 2019).