Addressing the tacit knowledge of a digital library system

Angela Di Iorio Marco Schaerf

DIAG - Department of Computer, Control, and Management Engineering Antonio Ruberti
Sapienza University of Rome, Italy

Abstract

Recent surveys, about the Linked Data initiatives in library organizations, report the experimental nature of related projects and the difficulty in re-using data to provide improvements of library services. This paper presents an approach for managing data and its "tacit" organizational knowledge, as the originating data context, improving the interpretation of data meaning. By analyzing a Digital Libray system, we prototyped a method for turning data management into a "semantic data management", where local system knowledge is managed as a data, and natively foreseen as a Linked Data. Semantic data management aims to curates the correct consumers' understanding of Linked Datasets, driving to a proper re-use.

1 Introduction

Nowadays, the implementation of Semantic Web Technologies (SemWebTech) in an existing Information System (InfSys) of an Organization (ORG) can be considered almost a necessary evolution, because it allows to deal with the pervasiveness of technologies and with the huge amount of deriving data.
In a Digital Library (DigLib) as a long-standing ORG, where multi-media objects are managed from their acquisition through their entire digital life-cycle, data related to multi-media maintenance process is ever-growing. As a consequence, the data management practices are 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 is part of the SemWeb, and focuses on the way data should be provided and shared by ORGs for achieving the vision of SemWeb, where LD is supposed to be re-used by consumers (humans and machines) over web protocols.
According to the Linked Data Best Practices published by the World Wide Web Consortium [W3C14]:

> [...] capture the context of data [...] high quality of Linked Data is obtained since capturing **organizational knowledge** about the meaning of the data within the RDF 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.

we focus on the *organizational knowledge* as a key component driving the understanding of data, and in particular on the "tacit" knowledge that is not already made explicit by the existing SemWeb ontologies, as LD vocabularies (LOV) [VAPVV17] providing LD with its meaning.
As an example, by analyzing the data of a local existing DigLib system of the Sapienza University, we found that data even defined by existing ontologies expressing knowledge about well known DigLib metadata standards, the

In: Marco Schaerf, Massimo Mecella, Drozdova Viktoria Igorevna, Kalmykov Igor Anatolievich (eds.): Proceedings of REMS 2018 – Russian Federation & Europe Multidisciplinary Symposium on Computer Science and ICT, Stavropol – Dombay, Russia, 15–20 October 2018, published at http://ceur-ws.org

data management of the local system relies on a "tacit" [Pol66] *organizational knowledge* that should be made explicit, as an ontology, in order to support the consumers' understanding of the meaning of data.
In this paper we address the problem of managing data and its knowledge for providing quality LD, driving data management toward the "semantic data management".
Transforming data management practices into "semantic data management practices" requires 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.). This transformation implies to consider InfSyss' knowledge as a data in the data management practices.
Semantic data management curates the correct consumers' understanding and the correct interpretation of exhibited LD.
In this paper, we present two local ontologies obtained by analyzing the DigLib system and detecting the underlying "tacit" knowledge. We show how we have dealt with "tacit" knowledge capture, also in relation to existing ontologies, from different knowledge domains.
The matching between local ontologies and existing ontologies drives the production of LD datasets, whose meaning is supported also by the *organizational knowledge*, as it is mentioned by the LD best practices, and it is considered essential for the correct interpretation of LD. The remainder of this paper is structured as follows. Section 2 reports the state of the art of LD implementation in ORGs managing DigLibs. Section 3 provides an explanation of the semantic data management. Section 4 explains the knowledge types. Section 5 overviews the explicit knowledge of the DigLib system case study. Section 6 describes the method for classifying the "tacit" knowledge of the DigLib system. Section 7 presents resulting local ontologies as a LOV. Section 8 draws the conclusions and presents the future developments.

2 State Of the Art

The 2nd Survey Report of the On-line Computer Library Center (OCLC) [1] published by Smith-Yoshimura in 2016 [SY16] reports the analysis of 112 Linked Data projects or services undertaken by 90 institutions in 20 different countries. The analysis, performed in 2015, describes the respondents' motivations as publishers or consumers of LD and indicates that most of the initiatives are primarily experimental in nature.
Among the most mentioned motivations we report the most relevant to this paper: *(5)* the need to publish linked data to consume it and to re-use it in future projects; *(6)* maximize interoperability and reusability of the data; *(7)* provide stable, integrated, normalized data on research activities across the institution.
These motivations highlight the specific interest of disseminating and re-using data, that implies to consider not only the perspective of the end-user as a consumer, but also the perspective of the ORG as a data manager and provider that exposes LD.
The recent survey of Tosaka and Park [TP18] 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. The survey identifies specific knowledge gaps to be addressed by data repository systems and specifically in relation to SemWebTech. The survey still reaches the conclusion of the exploratory stage of the LD implementation.

3 The Semantic Data Management

The management of data, and in turn the management of information and knowledge is one of the most studied and developed field in the automation of information, which nowadays is challenged by the automation of "Semantics" conveyed by the hierarchy of data, information, knowledge and wisdom [Row07].
The main difference between data management and semantic data management is that, in data management the semantic context of data is managed by persons affiliated to the ORG, by means of their human information, implicit and explicit knowledge, wisdom [Row07], while in the semantic data management, the data is equipped with its semantic context, which is codified in a machine-interpretable form (SemWebTech). The semantic data management provides machines with data and its knowledge context in a SemWebTech form, and the interpretation can benefit both machines and humans. Machines might re-use or re-manage data in a proper way, supported by knowledge driving the understanding of "why data has value", humans might be unloaded by long and discontinue searches of additional information for understanding of "why data has value".
Nevertheless capturing relevant knowledge context for data is challenging. The literature is rich of work generating ontologies from data and metadata of relational databases, but the underlying knowledge not always

[1] http://www.oclc.org/research.html

is explicit and is scattered into software documentation (also not always well-documented) or into technical reports, and its retrieval often is a time-consuming task.
The "tacit" knowledge, inadvertently implicit, hidden and given into the data management practices, is essential for enabling machines to properly interpret data, thus its detection, and its codification as a LOV, is an essential part of the semantic data management.
The adoption of SemWebTech in an existing data management system, implies indeed to understand how data can be interpreted by a third party, and as such, how the "explicit" and "tacit" knowledge about data management systems, has to be captured and provided as the data context.

4 Tacit Knowledge

The ORG knowledge is a research field developed since early nineties [Wii94]. Taking into account the most cited work of the field, comprehensively described by Evans et al. [EDB15], we review some theoretical analysis of the field supporting the method experimented for managing and structuring the ORG knowledge.
The ORG knowledge, as the data context, should be comprehensive of explicit and "tacit" [Pol66] [Gra96] knowledge.
The explicit knowledge in the knowledge management literature is distinguished in:

- "*Codified*" knowledge or knowledge that can be stored or put down in writing without incurring irreparable losses of information [Cho96]. This form of knowledge is highly refined [Wii94] and formalized, which allows it to be disseminated, more easily, more rapidly, and more extensively in the ORG than other forms.
- "*Encapsulated*" knowledge, which is not fully codified, and it is object-based, since the substantive knowledge that went into the design and development of artifacts remains partially hidden from its users [vdB13]. This is exactly the case of the RDB, where substantial expertise has been spent in its design as well as in its data population.

  > Encapsulation consists of the transformation of substantive knowledge into a product that requires only functional knowledge for its utility [vdB13]. Extracting and codifying encapsulated forms of knowledge requires further unpacking using methods similar to reverse engineering or compositional analysis.

  However, encapsulated knowledge is difficult to be collected and may also be subject to a misappropriation.

The "*tacit*" knowledge definitely is the source of the "codified" and "encapsulated" [vdB13] knowledge, which provides the "grounding of meaning" and the basis for the interpretation of a tacit activity [EDB15].

5 The Explicit Knowledge of a Digital Library System

Figure 1 shows the architecture of the Sapienza Digital Library InfSys [CDS14] where a Digital Resource (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) [Con12]. Conforming with the reference standard OAIS, the SDL DigRes is used for Submission (SIP), Archival (AIP) and Dissemination (DIP) functions, performed by DigLib application systems.
On the left of the Figure 1 different application systems perform the Dissemination function, as an example the SDL-WEB portal. On the right two specific applications, the Massive Conversion system and the Cataloguing System, performing the Submission function, create the DigRess conforming with requirements of the SDL InfSys. Specifically, the Massive Conversion (MassConv) DigLib system is the data management case study.
The MassConv is a data management system based on a relational database, where data managed are used for creating SDL DigRess conforming with well-established DigLib metadata standards. The MassConv manages data, describing the SDL multimedia objects. An automatic process extracts data from MassConv and creates XML files associated, by URI reference, with the multimedia objects. This process allows to build the DigRess composed by XML metadata files and multimedia objects.
The MassConv validates, the XML files created, against the XML schemas of the following standards:

- the Metadata Objects Description Schema [2] (MODS), for describing the intellectual contents of multimedia objects;

[2] Metadata Encoding Transmission Schema, `www.loc.gov/standards/mods/`

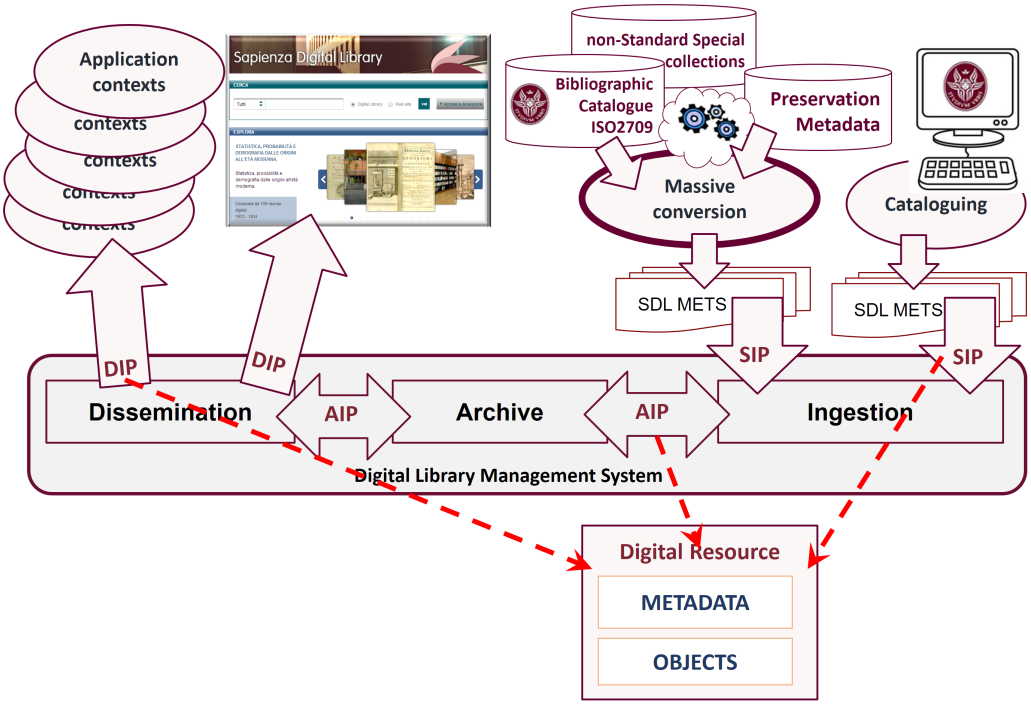

Figure 1: The SDL Digital Resource in the context of the DigLib-MS and related to the different OAIS IP types: submission, archival, dissemination

- the PREservation Metadata Implementation Strategies (PREMIS) [PRE15] for managing preservation metadata about multimedia objects;
- the Metadata Encoding and Transmission Standard (METS)[3] for DigRes packaging.

For the purpose of this article we focus on MODS and PREMIS, as a sample of explicit knowledge managed by the MassConv system. By analyzing the data content, of such explicit knowledge, we observe that data can be used by other relevant ontologies from other knowledge domain, and the exhibition of these knowledge connections as LOV allows data to be interpreted also by consumers, coming from other knowledge domain. Thus in the following subsection we briefly present the ontologies strictly representing the MassConv explicit knowledge (MODS and PREMIS) and the ontologies extending the interpretability of data, the Organization Ontology (ORG-O) and the Provenance Ontology (PROV-O).

5.1 The "Codified" Knowledge Matching with Different Domains

The explicit knowledge about data, managed by the local system, is formally codified in the following LOVs [VAPVV17]:

- **Metadata Object Description Ontology (MODS-O)** MODS-O 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 models the knowledge domain of digital preservation metadata, and develops around four main classes:
  `premis:Object`, `premis:Event`, `premis:Agent` and `premis:Rights`.
- **Provenance Ontology (PROV-O)** PROV-O [W3C13b] describes the concepts related to the provenance in heterogeneous environments, and develops around three main classes: Agent, Entity, and Activity.
- **Organization Ontology (ORG-O)** ORG-O [W3C13a] 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.[...]".

[3] Metadata Encoding Transmission Standard, `www.loc.gov/standards/mets/`

6 Detecting Tacit Knowledge in the Local System

By analyzing the MassConv database, we have realized that the explicit knowledge stored into the RDBs (data and schema) contains part of the system knowledge, that has driven its development. Commonly, the timeliness and costs plays the main role for reaching observable results. This fact drives worker staffs to neglect the capture of knowledge, produced during the period of system development. During the MassConv development, most of the knowledge locally "created and used" (see de Vasconcelos [dVKCR17] mapping between Knowledge Life Cycle and Software Development Life Cycle), remains scattered in text documentation which is difficult to be retrieved, and to be systematized, thus the knowledge context of data cannot support the understandability of data. This problem is unavoidably inherited by processes dealing with the generation of LD from the MassConv RDB.
Thus we have adopted the method of *a)* collecting software functions parameters passed through the massive conversion workflow; *b)* matching written definitions in the text documentation; *c)* creating identifiers for parameters as piece of embedded knowledge; *d)* creating a local ontology as the knowledge artifact for turning "tacit>embedded" knowledge into explicit knowledge; *e)* matching local ontology to existing ones. Consequently, tacit knowledge was captured and formalized in a local ontology, expressing the knowledge underlying the MassConv system.

6.1 Ontology for Software Embedded Knowledge

The implicit founding concepts for the management of SDL InfSys digital assets, were "codified" into a local ontology, named On-SDL. The main classes of the ontology are:

- Organizational Collection (OrgColl)
  The Organizational Collection provides an abstraction layer documenting the evolving history of the physical ORGs (`premis:Agent`), dealing with different legal aspects (`premis:Right`), and its changes (`premis:Event`) involving the maintenance of the Digital Objects (`premis:Object`).
- Digital Collection (DigColl)
  The Digital Collection is a special type of Digital Resource (DigRes) that collects data, inherited by the belonging DigRess, and it is based on the collecting activity of the ORG as a DigRes producer or maintainer. Data collected documents the production workflow.
- Digital Resource (DigRes)
  The Digital Resource is the "simplest" set of information coherently managed by a SDL system describing an Intellectual Entity [PRE15] conforming with the SDL metadata profile. The DigRes is the virtual set of Digital Metadata Objects (DMOs) and Digital Content Objects (DCOs).
- Digital Metadata Object (DMO)
  The Digital Metadata Object is a text file of whatever format (XML, CSV, JSON, RDF) comprehending data and metadata describing `premis:Objects` managed by the DigLib system.
- Digital Content Object (DCO)
  The Digital Content Object is whatever resource, (a multimedia file, a database...) which needs to be managed by the SDL system.

These classes are the parameters most used by the software functions, composing the MassConv system.
The ontology formalization for representing the main concepts, roles and individuals, of On-SDL, have been initially expressed in $\mathcal{ALC}$ the basic DescLogs [NB+03].
A DescLogs Ontology $\mathcal{O}$ consists of a TBox $\mathcal{T}$, and an Abox $\mathcal{A}$, respectively representing the intentional and the extensional knowledge [Baa03].
Figure 2 depicts the TBox $\mathcal{T}$ modelling the intentional knowledge, managed by the MassConv software:

- UniversityORG `subClassOf` UniversityDL: Sapienza organizational structure is reproduced in the SDL.
- UniversityORG `subClassOf` prov:Agent: Sapienza ORG are type of PROV-O agents.
- org:Organization `equivalent` UniversityORG: Sapienza is a type of ORG.

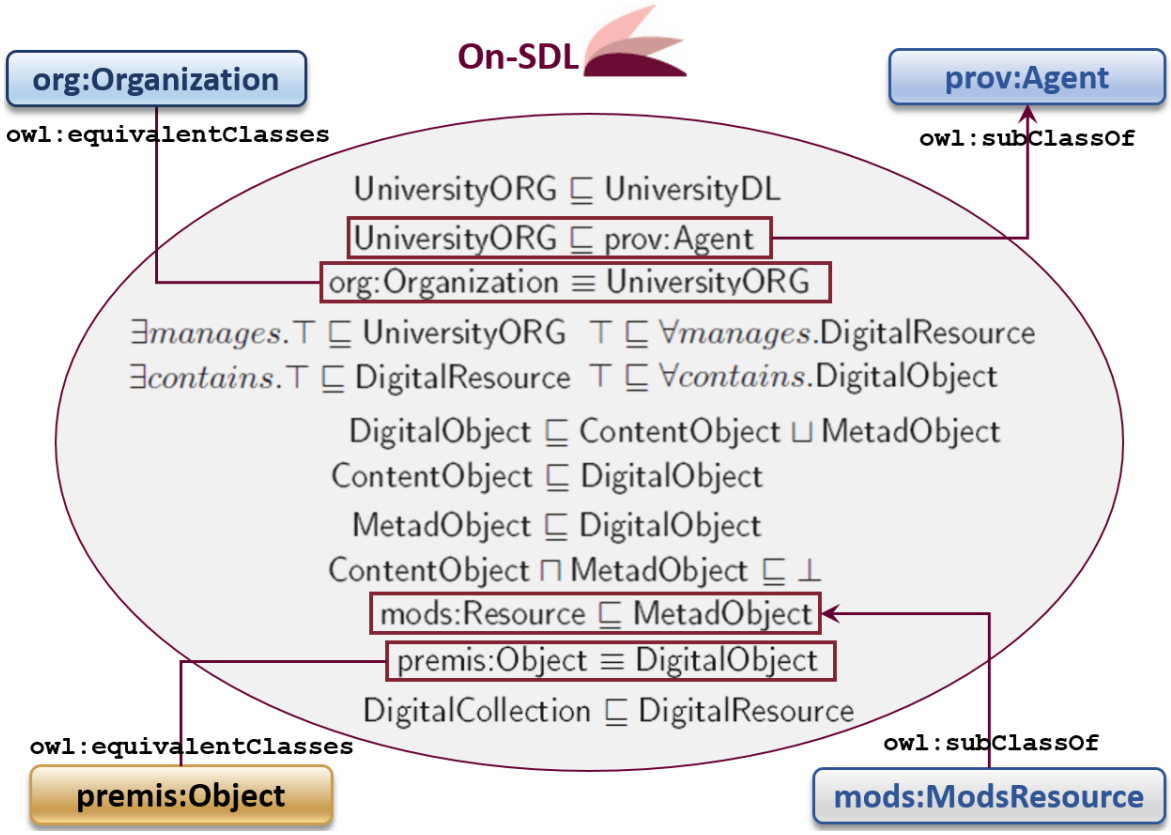

Figure 2: The TBox for the local system representing the tacit knowledge, and the matching with existing ontologies

- DigitalObject `subClassOf` ContentObject ⊔ MetadObject: DigObj is either a DCO or a DMO.
- ContentObject `subClassOf` DigitalObject: DCO is a type of DigObj.
- MetadObject `subClassOf` DigitalObject: DMO is a type of DigObj.
- ContentObject ⊓ MetadObject `subClassOf` ⊥: Nothing can be both DCO and DMO.
- mods:ModsResource `subClassOf` MetadObject: MODS Resource is a type of DMO.
- premis:Object `equivalent` DigitalObject: PREMIS Object is equivalent to DigObj.
- DigitalCollection `subClassOf` DigitalResource: DigColl is a type of DigRes.

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

- 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.

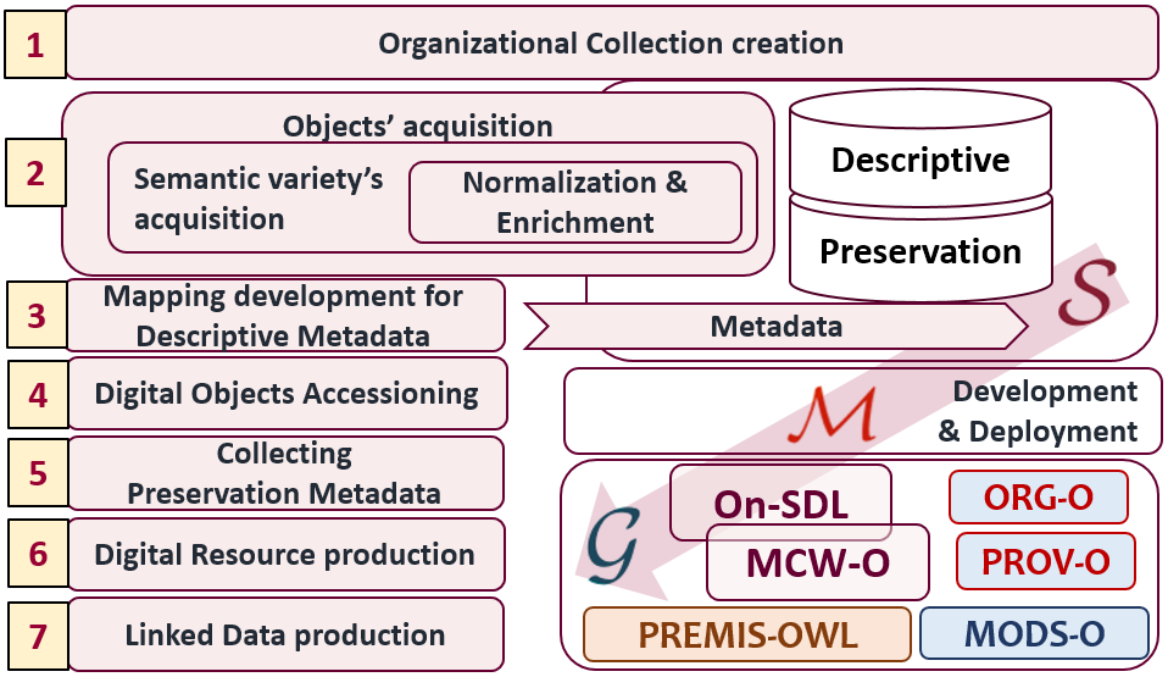

Figure 3: Workflow steps, and knowledge modeling from data sources to Linked Data Vocabularies

6.2 Ontology for Workflow Embedded Knowledge

MassConv was developed for managing the DigRes production workflow, based on Information integration Global-As-View approach [Len02], where the RDB is the data management technology (the data source $\mathcal{S}$), that is mapped by $\mathcal{M}$, toward global schema or ontology $\mathcal{G}$. Figure 3 depicts on the right this process and on the left, the MassConv workflow steps, that are described as follows:

1. *Organizational Collection creation* identifies (with a root URI) a Sapienza ORG.

2. *Object Acquisition* stores DCOs from a Sapienza ORG, into a working area, assigns to the DigRess, the URI based on the ORG identifier, associates the related descriptive data (MODS), and computes or collects preservation data (PREMIS).

3. *Mapping Development* builds the conversion layer toward MODS and PREMIS semantics.

4. *Object Accessioning* stores DCOs in the SDL repository, from the *Acquisition* working area, propagates and extends DigRess' URIs over belonging DCOs.

5. *Collecting Preservation Metadata* about DCOs is automatically gathered and computed.

6. *Digital Resource production* DMOs and related DCOs are produced, according to the SDL XML metadata schemas $\mathcal{G}$.

7. *Linked Data production*, the last step to be developed, re-uses data at the data source $\mathcal{S}$, and extends the mapping $\mathcal{M}$ necessary to local ontologies $\mathcal{G}$.

We can observe that each workflow step is knowledge, embedded in the software functions performing each step. Consequently, that knowledge was codified into a local ontology for describing the MassConv Workflow (MCW-O).
Figure 3 shows on the right, how the "tacit" knowledge, that we have codified, flows from the data source $\mathcal{S}$ toward the On-SDL and the MCW-O, and in turn routes data toward existing LOVs, the PREMIS-OWL, the MODS-O, the PROV-O, the ORG-O.

7 Tacit Knowledge Codified as a Linked Data Vocabulary

Figure 4 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. The MCW-O classes can be distinguished by the yellow tags, numbered according to the workflow steps, that are described above.
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 SDL 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. This demonstrates that the requirement of the LD principles prescribing to adopt LOV is fulfilled.
The second LD requirement for using URIs, is already covered by the MassConv system, as witnessed by the workflow itself. As a DigLib system producing DigRess to be exchanged among different systems performing OAIS functions, the MassConv system was natively equipped with an URI management method [DIS14]. In the Fig. 4, it is possible to see the specific events described as workflow steps (*1*, *2* and *4*), that manage the URI. Thus, LD URIs will be generated by extending the method to the entities detected for local ontologies and to data itself.
Both data and vocabularies expressing the local organizational knowledge can be then generated as Linked Datasets.

8 Limitations, Conclusions and Future Developments

The method adopted for capturing tacit knowledge from the DigLib system case study is a prototype, that should be evaluated in other InfSyss. The detection of the tacit knowledge is in any case not a straightforward task, mostly is manual and requires the effort of the knowledge worker for being performed.
Nevertheless, the method can be a training for developing a mindset of knowledge workers, oriented toward the management of LD pillar elements, the URI and LOV. The transition into "semantic data management", is based on the management of the system knowledge, likewise of the system data: knowledge data is codified in a "machine-interpretable" form. The computable interpretation of data allows machines to support better the human work in understanding of "why data has value", and thus in re-using or re-managing data in a proper way.
In the near future, we will be developing the workflow for generating the foreseen Linked Dataset, referring to the described On-SDL and MCW-O ontologies.

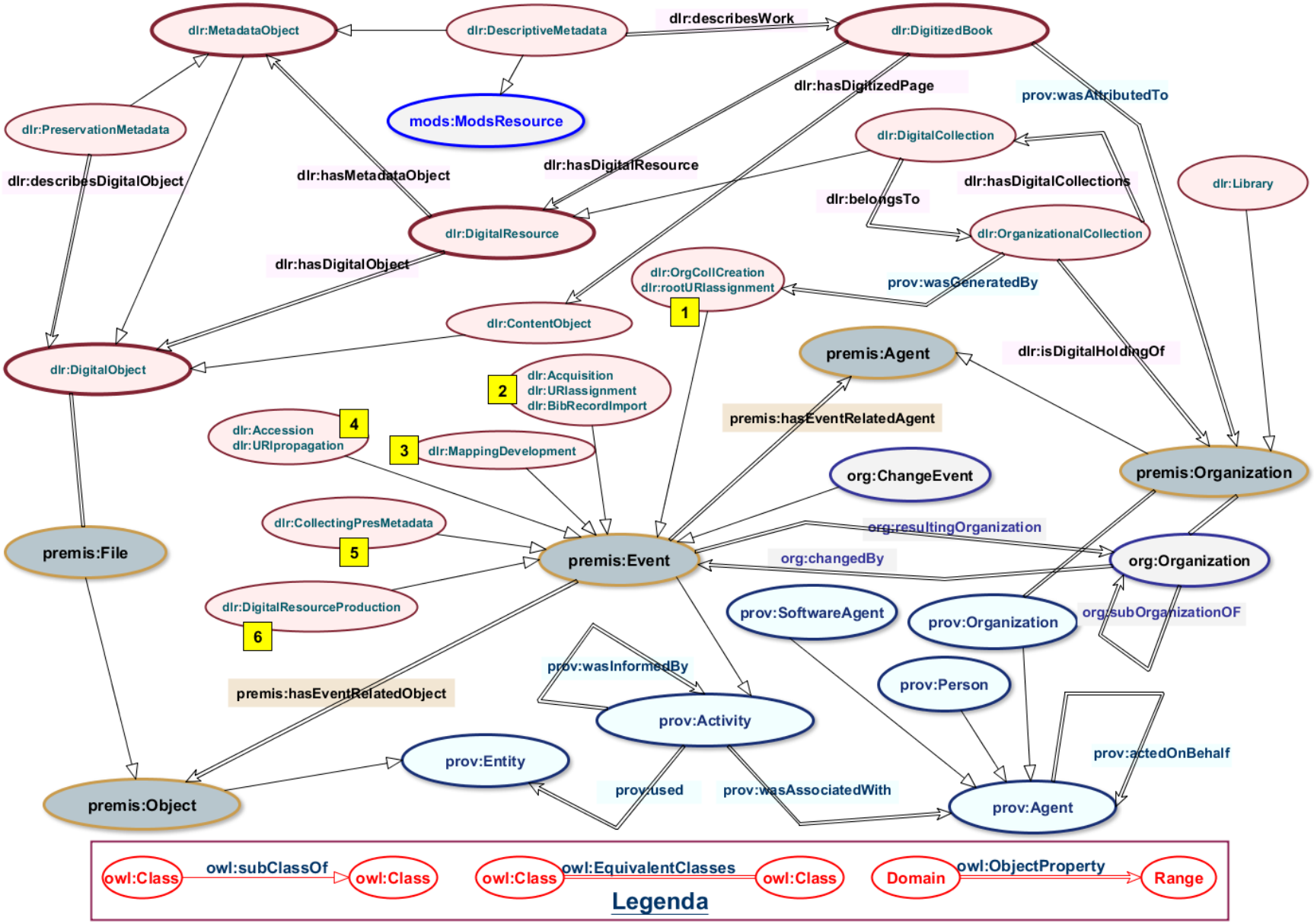

Figure 4: Workflow steps in the context of ontologies SDL-O and MCW-O, related to PREMIS-OWL, PROV-O, ORG-O, MODS-O

References

[Baa03] Franz Baader. *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[CDS14] Tiziana Catarci, Angela Di Iorio, and Marco Schaerf. The sapienza digital library from the holistic vision to the actual implementation. *Procedia Computer Science*, 38:4–11, 2014.

[Cho96] Chun Wei Choo. The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. *International journal of information management*, 16(5):329–340, 1996.

[Con12] Consultative Committee for Space Data. Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book, 2012.

[DIS14] A Di Iorio and M Schaerf. Identification semantics for an organization, establishing a digital library system. *Semantic Digital Archives*, page 16, 2014.

[dVKCR17] José Braga de Vasconcelos, Chris Kimble, Paulo Carreteiro, and Álvaro Rocha. The application of knowledge management to software evolution. *International Journal of Information Management*, 37(1):1499–1506, 2017.

[EDB15] Max Evans, Kimiz Dalkir, and Catalin Bidian. A holistic view of the knowledge life cycle: the knowledge management cycle (kmc) model. *The Electronic Journal of Knowledge Management*, 12(1):47, 2015.

[Gra96] Robert M Grant. Toward a knowledge-based theory of the firm. *Strategic management journal*, 17(S2):109–122, 1996.

[Len02] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA, 2002. ACM.

[NB+03] Daniele Nardi, Ronald J Brachman, et al. An introduction to description logics. In *Description logic handbook*, pages 1–40, 2003.

[Pol66] Michael Polanyi. *The Tacit Dimension*. 1966.

[PRE15] PREMIS Editorial Committee. PREMIS Data Dictionary for Preservation Metadata, Version 3.0, 2015.

[Row07] Jennifer E Rowley. The wisdom hierarchy: representations of the dikw hierarchy. *Journal of information science*, 2007.

[SY16] Karen Smith-Yoshimura. Analysis of international linked data survey for implementers. *D-Lib Magazine*, 22(7/8), 2016.

[TP18] Yuji Tosaka and Jung-ran Park. Continuing education in new standards and technologies for the organization of data and information. *Library Resources & Technical Services*, 62(1):4–15, 2018.

[VAPVV17] Pierre-Yves Vandenbussche, Ghislain A Atemezing, María Poveda-Villalón, and Bernard Vatant. Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web. *Semantic Web*, 8(3):437–452, 2017.

[vdB13] Herman A van den Berg. Three shapes of organisational knowledge. *Journal of Knowledge Management*, 17(2):159–174, 2013.

[W3C13a] W3C. ORG-O: The Organization Ontology, 2013.

[W3C13b] W3C. PROV-O: The PROV Ontology, 2013.

[W3C14] W3C. Best Practices for Publishing Linked Data, 2014.

[Wii94] Karl M Wiig. *Knowledge Management Foundations: Thinking about Thinking-how People and Organizations Represent, Create, and Use Knowledge*. Schema Press, Limited, 1994.