Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 5th, 2017

Merging Enterprise Metadata for Minimizing Business Process Changes with Knowledge Graphs

LLiver José

Seidenberg School of CSIS, Pace University, Pleasantville, New York
jl71224w@pace.edu

Abstract— Researchers in the ontology-design field have developed the content and solutions for ontologies covering many overlapping domain areas[1]. Ontology alignment and merging have evolved from been usually handled manually - often constituting a large portion of the sharing process and artifacts of a tedious knowledge engineering, to become increasingly common on the World-Wide Web x.0, where they provide semantics for annotations in Web pages, Folksonomy, Business Data Management and Governance, etc. In order for these ontologies to be re-used, they first need to be merged or aligned; however, many complexities and issues arise when merging metadata for horizontal organizations. This work proposes a framework based on the outcome of several works to: Support ontology custom relations and data dictionary; Introduce mapping of abstract fields as the common base concept to align repositories; Automate the mapping of metadata fields to minimize the business process changes. These will allow better abstract metadata translation timeframe of application integration, better support for future merger, reduce domain experts' intervention.

***Index Terms*—Content Management, Metadata Management, Business Metadata, Business Process, Ontology, Natural Language Processing, Lexemes, Lexical Services, Parser.**

I. INTRODUCTION

Business Metadata describes various facets of an information asset in order to improve its usability throughout its life cycle. The growing need for organizations of all types to treat information as an asset is making metadata management strategic, driving significant growth for metadata management solutions. "Through 2018, 80% of data lakes will not include effective metadata management capabilities, making them inefficient. By 2020, 50% of information governance initiatives will be enacted with policies based on metadata alone."[2]

When two horizontal organizations (within similar sectors and line of business) join through a merger or acquisition (M&A) - unity requires merging or aligning both enterprises onto the same data platform in order to have a consistent view of the newly forged organization[3, 4]. In a merger, business process changes must be minimized to ensure data flow and decision making across disparate metadata are expeditious, accurate, globally strategic, unified and cost minimized. From these disjointed datasets, several challenges arise when data consumers need to extract global business intelligence and enforce a consistent unified representation for a) metadata complexity and volume, b) real-time visualization and c) apply consistent rules all throughout the business.

II. PRIMARY SUPPORTING WORKS

Improving Data Governance in Large Organizations through Ontology and Linked Data[4]. This research demonstrates that given the dynamic and complex nature of global organizations today, governing corporate data is a challenge; and proposes methods and technologies that can contribute to effective data governance in large organizations improving data awareness and governance. Reusing this work will assist me demonstrating how graph can be an effective manner to visualize data governance. My contribution here is to apply it to multi horizontal organizations merging their metadata sets instead of a single organization using KG instead of RDF.

Knowledge Graph Syntax Validation and Visual Navigation for Developing Intelligent Systems[5]. This research demonstrates the challenges that need to be overcome in-order to better develop a new algorithm to support Knowledge Graph (KG) syntax validation so that domain experts can develop valid and robust KGs and to support visual navigation to ascertain their completeness and logical accuracy. My contribution to this work is to re-use the KG work and apply it to horizontal merging organizations to minimize business processes.

Reducing Complexity of Diagnostic Message Pattern Specification and Recognition with Semantic Techniques[6]. This research demonstrate that data mapping is the glue that tie together information from various sources enabling the integration of information. Using XML Dialect Proliferation alone is an immense problem for Accurate Data Interchange.

XML files must adhere to a Published Contract, typically called a Standard. This current work uses extensively the findings of this dissertation to avoid the need of duplication of efforts. My contribution to this work is to extend it to use multiple sources of document types as well as using API for database data extraction.

Algorithm and Tool for Automated Ontology Merging and Alignment[1]. This research developed and implemented PROMPT, an algorithm that provides a semi-automatic approach to ontology merging and alignment. PROMPT performs some tasks automatically and guides the user in performing other tasks for which his intervention is required. PROMPT also determines possible inconsistencies in the state of the ontology, which result from the user's actions, and suggests ways to remedy these inconsistencies. This is based on an extremely general knowledge model and therefore can be applied across various platforms. The formative evaluation showed that a human expert followed >90% of the suggestions that PROMPT generated and that >74% of the total knowledge-based operations invoked by the user were suggested by PROMPT. This is the backbone algorithm used within the framework of this solution for data automation.

III. Use cases: Business Metadata Merging / Aligning

Merging is the process of finding commonalities between two different ontologies A and B and deriving a new ontology C that facilitates interoperability between horizontal organizations that are based on the A and B ontologies. The new ontology C may *replace* A or B, or it may be used only as an *intermediary* between a system based on A and system based on B. Depending on the amount of change necessary to derive C from A and B, different levels of integration can be distinguished.

The following use cases presented below focus on mapping multiple metadata across horizontal organizations containing synonyms, homonyms, polysemy, similar terms with different properties, and abstract concepts of similar instances with different meanings and processes.

A. Use Case 1: Metadata mapping of abstract concepts refinement across two merging horizontal organizations.

The abstract concept attached to metadata term "Region" in Organization 1 means *Time Zone*; while in Organization 2 means *State*. In Organization 2 *Region* metadata definition is a subset of Organization 1 - because a Time Zone is comprised of States. *State* holds more granular detail of the true meaning in the newly formed metadata repository; therefore, a *concept refinement* is necessary, where *Region = State*, instead of Time Zone. Figure 1 shows the use case of mapping metadata concepts, the main focus of this paper.

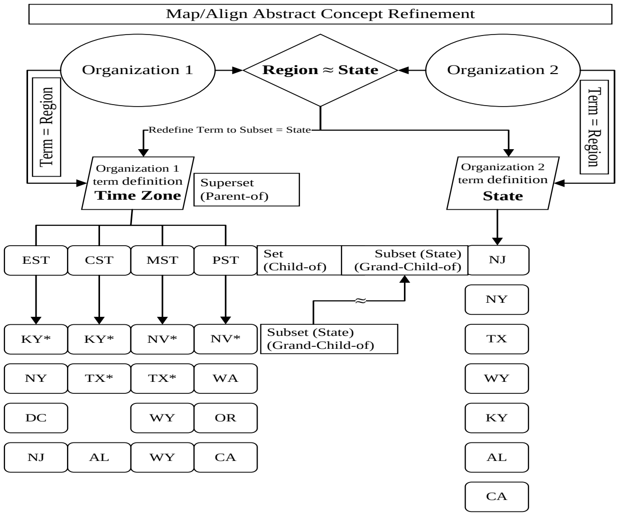
Map/Align Abstract Concept Refinement
Organization 1
Region ≈ State
Organization 2
Term = Region
Term = Region
Redefine Term to Subset = State
Organization 1 term definition Time Zone
Superset (Parent-of)
Organization 2 term definition State
EST
CST
MST
PST
Set (Child-of)
Subset (State) (Grand-Child-of)
NJ
≈
NY
KY*
KY*
NV*
NV*
Subset (State) (Grand-Child-of)
TX
NY
TX*
TX*
WA
WY
DC
WY
OR
KY
NJ
AL
WY
CA
AL
CA

Figure 1 Metadata Mapping Abstract Concept Refinement

B. Use case 2 – Mapping metadata synonyms, homographs, and polysemies across two horizontal organizations.

The following Figures *2-4* are sub-problem categories of metadata merging outside of the scope and can be an extension of this work. These use cases show the mapping of equivalent metadata terms, synonyms, and homonyms.

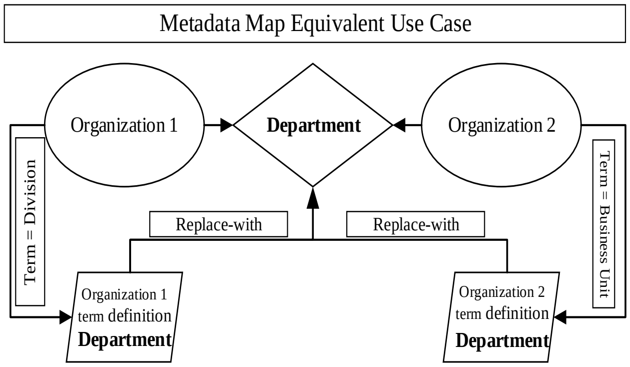
Metadata Map Equivalent Use Case
Organization 1
Department
Organization 2
Term = Division
Term = Business Unit
Replace-with
Replace-with
Organization 1 term definition Department
Organization 2 term definition Department

Figure 2 Metadata Map Equivalent Terms Use Case

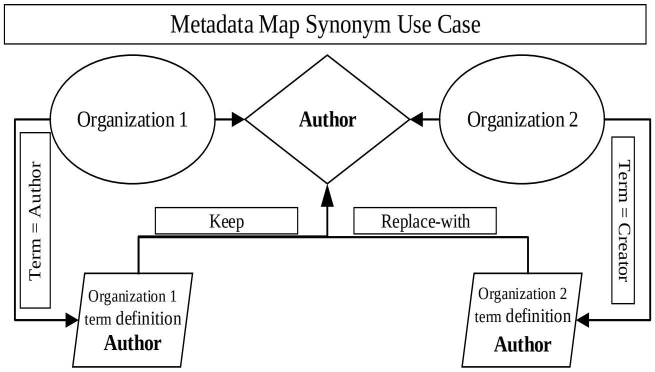
Metadata Map Synonym Use Case
Organization 1
Author
Organization 2
Term = Author
Term = Creator
Keep
Replace-with
Organization 1 term definition Author
Organization 2 term definition Author

Figure 3 Metadata Map Synonym Use Case

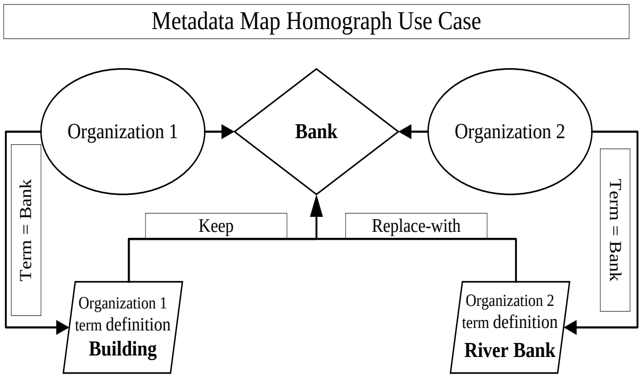

Figure 4 Metadata Map Homograph Use Case

C. *Use case 3 – Mapping terms specifications across two horizontal organizations.*

Figure 5 shows a special category of metadata merging wherein both organizations the metadata term is named and used similarly; however, one or more aspects can differ either in *data type, format* and/or *size*; its usage may result in data loss. This use case is a sub-problem also outside of the scope and can be an extension of this work.

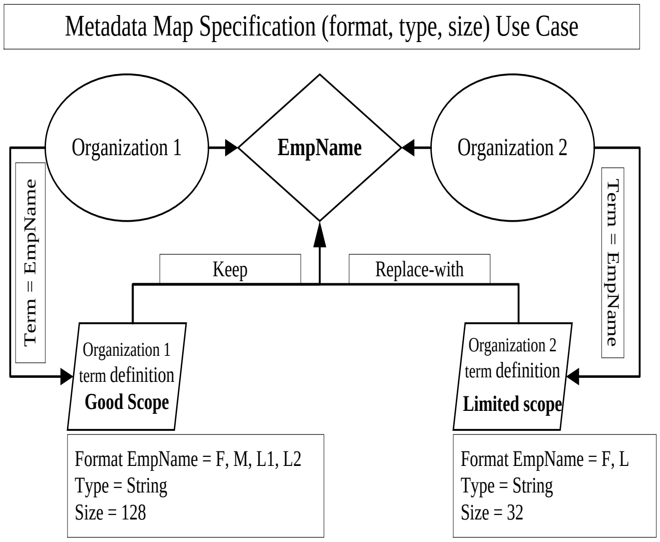

Figure 5 Metadata Map Terms Specification Use Case

IV. Importance of Merging Business Metadata

There are tangible benefits to streamline and minimize the processes interfacing with enterprise business metadata management via knowledge graph (*Pace University extended Protégé*, a tool for extensible knowledge representation supporting custom relations for domain experts to describe and validate knowledge[5].) For example, this helps data visualization identifying stakeholders, affected processes and minimize financial impacts which range in the billions of dollars of waste and loss of opportunity to businesses[4]. Enterprise business metadata is important that it be available, accurate, and timely in order to appropriately make decisions on managing corporate data assets and enterprise knowledge[7].

Having on-premise or a cloud hosted global repository *alignment* (sources are made coherent with one another but kept separately) or *merging/mapping* (single coherent ontology that includes information from all the sources) of the business metadata across the M&A will aid in identifying the system catalogs, affected stakeholders, interrelationships of systems and how they attach to processes that would be affected[4]. This also will increase: a) speed of data access; b) accurate results by having everyone communicating with the same semantic; c) global access; d) cost of access; e) business processes minimized; f) better application integration. Additional important considerations are:

a) Reduce time required to find a piece of information. Finding where to look is the biggest problem in understanding far-flung and disparate pieces of data especially when they are dispersed within new mergers.
b) Reusing data helps the organization to run more efficiently. When there is no metadata directory, the organization is faced with building everything from scratch every time a new request for information arrives.
c) The accuracy of information. Business analysts spend approx. 80% of their time gathering and validating data [7].

V. Current Related Works.

A. *Relevant Literature Review*

The following Table 1 shows some of the most relevant current solutions available to address metadata content management. These are about the organization's management of data and information assets to address use cases such as data governance, analytics and enterprise metadata management (EMM). It is important to note that this understanding of metadata goes far beyond just technical facets; it is used as a reference for business-oriented and technical projects and builds the foundations for information governance and analytics[2].

Company	**Solution & Main Strength or Weakness**
Adaptive	*Adaptive Metadata Manager.* Broad metadata management use cases: such as those for big data and analytics, or regulation and compliance — through a combination of products. Poor integration of products because users of metadata management tools are evolving toward business functions.
Cambridge Semantics	*Anzo Smart Data: Platform, Integration and Manager.* The semantics standards approach of Cambridge Semantics' provides great alignment to initiatives that are based on open linked data or World Wide Web Consortium (W3C) Semantic Web standards — most notably RDF/OWL. However, this approach requires additional layers of mapping, and therefore additional complexity, to support the fact that metadata is context-sensitive
Collibra	*Collibra Data Governance Center.* Data governance and information stewardship. Not centered around technical metadata management capabilities, only business metadata.
Data Advantage Group	*MetaCenter.* Maintains a library of hundreds of virtual machine images.

Global IDs	*Global IDs Enterprise Information Management (EIM) Suite and Global IDs Ecosystem Management Suite.* Approach to metadata management is based on machine-centric automated learning enabling automatic curation of metadata assets. The metadata repository is graph-oriented and can be composed of any number of subgraphs representing things such as data elements, applications or business concepts.
IBM	*InfoSphere Information Governance Catalog.* Depth and breadth of usage apply to a wide variety of data domains and use cases.
Informatica	*Metadata Manager, Business Glossary and Enterprise Information Catalog.* Alignment with evolving trends and business-facing demand such as enterprise data catalog, self-service data preparation, stewardship and governance, and data analytics — with the graph-based Live Data Map of metadata assets — represents a focus on maximizing business value.
Oracle	*Oracle Enterprise Metadata Management.* Synergies with portfolio's broad range of technologies.
SAP	*PowerDesigner and Information Steward.* Broad product portfolio, embracing many aspects of information governance such as data integration, master data management, data quality, information lifecycle management and metadata management.

Table 1 – Most Relevant Metadata Solutions Available

B. *Current Solutions Limitations*

Most of the current solutions address a more comprehensive line of the business problem involving technical and business metadata in general; instead of within M&A. There is not a single solution that stands out to efficiently solve the problem referring to "enterprise business metadata management for minimizing business processes using knowledge graph", and thereby the nature of this work. Some of the deficiencies observed in the current solutions are:

a) The semantics standards approach provides great alignment to initiatives that are based on open linked data or World Wide Web Consortium (W3C) Semantic Web standards — most notably RDF/OWL. However, this approach requires additional layers of mapping, and therefore additional complexity, to support the fact that metadata is context-sensitive[2].
b) The current solutions tend to have high costs, complex implementations, or a ridged structure associated with them, providing more barriers to adoption.
c) None of the solutions focus on providing metadata merging during the initial stage of the merger negotiation, prior to contract signing.
d) The "technology assumes that all systems are semantically homogeneous – i.e. that they will all use the same vocabulary" [7, 8].
e) High level of complexity due to multiple products used.
f) Use of traditional metadata ingestion and management.
g) Managing governance and risk across complex data landscapes will require strong rules management capabilities and domain experts' intervention.

VI. IDEA FOR IMPROVING THE CURRENT SOLUTIONS

The main quantifiable improvements are to Minimize business metadata changes between the two newly merged horizontal organizations by consolidating into a single repository both enterprises; Minimize business process changes by introducing automation and reducing the need for domain experts intervention, Minimize metadata translation during business operation by using same abstract vocabulary for both organizations, and Better support future business merges by presenting a framework to ease the mapping of business metadata. Other areas this work can potentially improve or influence are:

a) Formalize existing process and spot needed improvements.
b) Facilitate identification for automation for efficient process flow.
c) Increase productivity and decrease domain experts head count interfacing with the processes.

Figure 6 shows two typical horizontal enterprise metadata mapping between organizations. When merging, three general use cases are inferred from here - detailed in Section II above: a) mapping abstract concepts of similar instances with different meanings and processes; b) synonyms, homonyms, polysemy; and c) mapping similar metadata terms with different specifications across the two organizations.

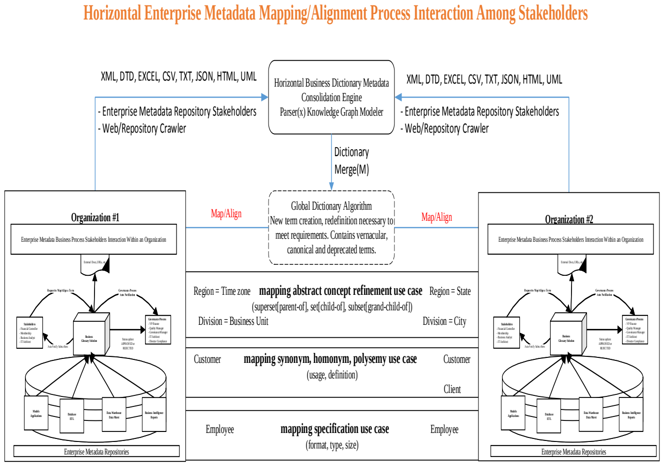
Horizontal Enterprise Metadata Mapping/Alignment Process Interaction Among Stakeholders
XML, DTD, EXCEL, CSV, TXT, JSON, HTML, UML
- Enterprise Metadata Repository Stakeholders
- Web/Repository Crawler
Horizontal Business Dictionary Metadata Consolidation Engine Parser(x) Knowledge Graph Modeler
XML, DTD, EXCEL, CSV, TXT, JSON, HTML, UML
- Enterprise Metadata Repository Stakeholders
- Web/Repository Crawler
Dictionary Merge(M)
Map/Align
Global Dictionary Algorithm New term creation, redefinition necessary to meet requirements. Contains vernacular, canonical and deprecated terms.
Map/Align
Organization #1
Enterprise Metadata Business Process Stakeholders Interaction Within an Organization
Enterprise Metadata Repositories
Organization #2
Enterprise Metadata Business Process Stakeholders Interaction Within an Organization
Enterprise Metadata Repositories
Region = Time zone
mapping abstract concept refinement use case
Region = State
(superset[parent-of], set[child-of], subset[grand-child-of])
Division = Business Unit
Division = City
Customer
mapping synonym, homonym, polysemy use case
Customer
(usage, definition)
Client
Employee
mapping specification use case
Employee
(format, type, size)

Figure 6 Horizontal Enterprise Metadata Mapping Between Organizations

Given two horizontal organizations with same purpose similar metadata representations in need of merging (reference use case 1):

1. Support relations (superset [grand-parent-of], set [parent-of], subset [child-of]), ontologies or knowledge graphs. Data dictionary (map tables) may be another proper tool.
2. Introduce abstract fields as the common base concept or the new field of the two existing ones (the new one could be one of the existing ones).
3. Define mapping rules of existing fields to the abstract ones.
4. Automate the mapping of metadata fields in the business process.

VII. PROBLEM STATEMENT

Organizations often merge in order to expand into new markets, acquire new talents, technology or reduce competition. This work looks at the abstract concepts subproblem within a larger metadata merging problem domain that includes synonym, homonym, polysemy, and metadata term properties specification. Important considerations for this problem domain are:

1. Multiple datasets across horizontal business boundaries must be merged to avoid maintaining similar copies relating to metadata synonyms, homonyms, and polysemy with similar or different meanings and processes, or in need of abstract terms refinement to accommodate the new merger, and optimize business processes approval - which are longer, costly and slow time to resolution because of data quality issues when kept separately.
2. Communication accuracy within the global organization—merging unity also requires bringing both enterprises onto the same data platform in order to have a consistent state of the newly forged organization[3, 4]. Enforce a consistent and global unified view for the complexity and volume of metadata, real-time visualization, and necessity to apply consistent business rules throughout the business.
3. The composition of different technologies, policies, cultures, repositories and data quality standards impact the overall merging of the business to provide the right information to the right stakeholders at the right time; having the right synergy between people, technologies, and processes[8].

A. *Solution Scope and Constrains*

The following are imposed constraints to the scope of this work:

1. This work applies to M&A of *horizontal enterprise business* metadata, where the two organizations are in similar business sectors and line of business.
2. Metadata merger applies only to *abstract concepts*; which is a sub-problem within a larger metadata merging problem domain that includes: synonym, homonym, polysemy and metadata field specification.
3. Organizations must have metadata repositories already in place.

However, this work does not apply to:

1. *Demerger*, spin-off, and spin-out. Where organizations split into two.
2. *Triangular merger*, forward triangular merger and reverse triangular merger, where the target company merges with a shell company, becoming a subsidiary in some form.
3. *Conglomerate merger*, where the two organizations are in irrelevant business.
4. *Vertical merger*, where the two organizations don't have core business similarities.
5. *Cartels*, where businesses secretly associate to maintain the profitability of the same good. Not suitable for this study because of the volatility and resources disjoint.
6. *Divestitures*, Equity Carve-outs, involves sell off of a portion of a firm to a third party. It involves cash transaction. Or,
7. *Quick merger* with similar business but unrelated technology and different management.
8. *Horizontal merger without metadata repositories*.

B. *Solution Objectives*

The objective is to merge horizontal enterprise metadata for minimizing business process changes with knowledge graphs. To solve the problem of disjointed vocabularies, we need to look at semantic-based information interoperability solutions. These solutions have three major characteristics[9]:

1. *Semantic Mediation*: this solution uses an ontology model that makes concepts explicit – as a mediation layer in order to abstract particular data terms, vocabularies, and information into a shareable and distributable model.
2. *Semantic Mapping*: mapping to an ontology maintaining the semantics of data and excludes the need for custom code.
3. *Context Sensitivity*: any information interoperability solution set must accommodate the fact that the same data can mean many different things from different viewpoints. Typically, the business rules, context definitions, and environmental metadata are captured and stored during the mapping process, making them reusable in any runtime server process.

The objective is to focus on merging enterprise metadata for minimizing business process changes with knowledge graphs to facilitate automation, reduce translation error, and increase visualization to the data flow. Figure 7 shows a typical well-defined enterprise with metadata processes and stakeholders.

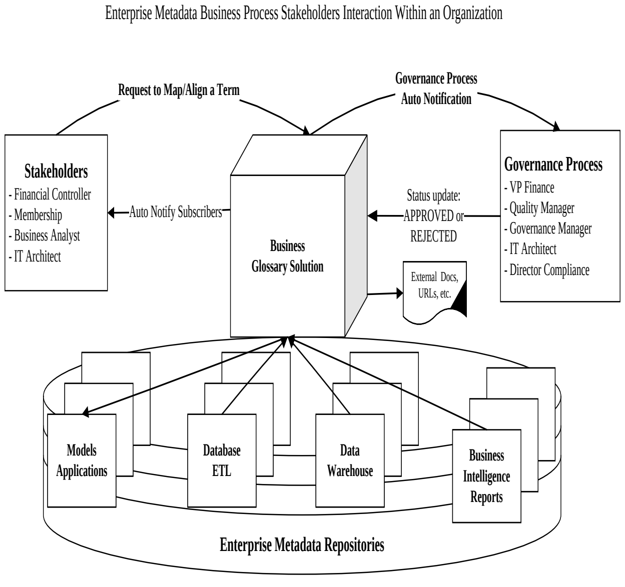
Enterprise Metadata Business Process Stakeholders Interaction Within an Organization
Request to Map/Align a Term
Governance Process Auto Notification
Stakeholders
- Financial Controller
- Membership
- Business Analyst
- IT Architect
Auto Notify Subscribers
Business Glossary Solution
Status update: APPROVED or REJECTED
Governance Process
- VP Finance
- Quality Manager
- Governance Manager
- IT Architect
- Director Compliance
External Docs, URLs, etc.
Models Applications
Database ETL
Data Warehouse
Business Intelligence Reports
Enterprise Metadata Repositories

Figure 7 Workflow to Map/Align Enterprise Metadata Repository

C. Tools and Techniques Used

1. Ontology and Visualizers / Query Tools
   - Protégé - To Create Knowledge Graphs

2. API's
   - Apache Jenna - Used to Parse KGs and Extract Objects and Relations

3. Web Search
   - Web Crawler – automatically and unattended crawl thru all repositories (CRM, DB, Web Pages, Document files) to build lexicon dictionary.

4. Semantic Validation
   - Schematron - Use Schematron Semantic Validation

5. PROMPT Algorithm
   - Merge and Align Repositories

VIII. General Solution Algorithm

This solution design algorithm framework shown in Figure 8 covers four primary areas: *Exception Handler*, where all exceptions take place; *User Interface*, where any interaction with the enterprise repositories and stakeholders are handled; *Parser and Domain Thesaurus*, where terms are analyzed; and *Lexical Service*, where a common dictionary and thesaurus of terms is created and maintained for merging/mapping or alignment.

This solution minimizes business processes by presenting a framework to automating most of the tedious matching manual work and removing most of the domain experts' interaction during the decision process; except in special occasions where they are absolutely necessary.

IX. Contribution and Conclusion

In harnessing data for business outcomes, data leaders must understand the flood of data in multiple formats. Information has been available in disparate repositories for decades, but in today's digital business environment organizations face new demands to access and use data across these repositories (especially when two horizontal organizations merge or align their respective repositories) — by mapping the relationships between different data elements. Reducing metadata business processes expedite availability and promote timely business intelligence as an ultimate goal of the competition.

Here we clearly showed that organizations will need to prepare, adjust to and exploit the following upcoming changes[2]:

1. The variety and extent of metadata supported across merging repositories.

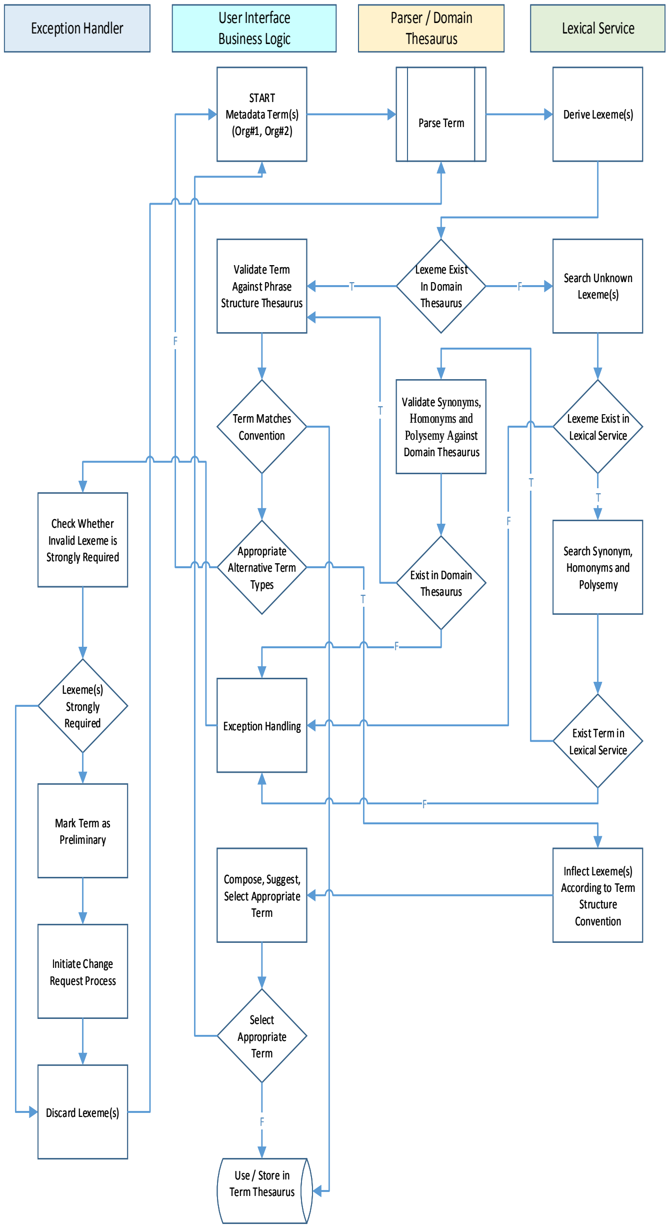
Exception Handler
User Interface Business Logic
Parser / Domain Thesaurus
Lexical Service
START Metadata Term(s) (Org#1, Org#2)
Parse Term
Derive Lexeme(s)
Validate Term Against Phrase Structure Thesaurus
Lexeme Exist In Domain Thesaurus
Search Unknown Lexeme(s)
T
F
Term Matches Convention
Validate Synonyms, Homonyms and Polysemy Against Domain Thesaurus
Lexeme Exist in Lexical Service
Check Whether Invalid Lexeme is Strongly Required
Appropriate Alternative Term Types
Exist in Domain Thesaurus
Search Synonym, Homonyms and Polysemy
Lexeme(s) Strongly Required
Exception Handling
Exist Term in Lexical Service
Mark Term as Preliminary
Compose, Suggest, Select Appropriate Term
Inflect Lexeme(s) According to Term Structure Convention
Initiate Change Request Process
Select Appropriate Term
Discard Lexeme(s)
Use / Store in Term Thesaurus

Figure 8 – General Solution Algorithm

2. The enhancement of the scope of metadata through automation (machine-learning, ontology, etc.), and through automated enrichment by semantic search capabilities, standard processes, and crowdsourcing.
3. Semantics formalism for improved interoperability between business processes, applications, and accuracy within mergers.
4. New ways to visualize metadata across a federated environment (self-service data preparation for analytics is driving this requirement)
5. New governance models, driven by the IoT and increased regulatory scrutiny.
6. The rapid transfer of metadata ownership from the CIO to the CDO is crucial maintain market leadership.

7. New technology innovations will generate interest in bridging information silos in order to improve the value of information-based business outcomes, for example: the ability to address a variety of data types and to capture and enrich metadata at the time it is being loaded; the ability to combine machine-learning and crowdsourcing metadata from experts; and support for complex environments to provide end-to-end data lineage. However, there are still a few inhibitors to even faster adoption, including:
   a) The lack of maturity of strategic business conversations about metadata
   b) The required and expensive effort of integration for metadata management solutions in multi merging horizontal environments.
   c) The lack of identification of accurate metadata management solutions whose capabilities match the current and future requirements for specific use cases.

Most organizations will find that their current metadata management practices are different across applications, data, and technologies and that these practices are siloed by the needs of different disciplines — each with their own governance authority, practices and capabilities. Data and analytics leaders that have already invested in data management tools/solutions should first evaluate the metadata management capabilities of their existing data management tools, including their federation/integration capabilities, before investing on a new metadata management solution. However, if they are dealing with emerging use cases — including collaborative analytics and community-oriented data governance — they should learn about, and investigate new metadata management solutions.

This work can be extended by focusing on metadata merging horizontal organizations synonyms, homonyms, polysemies, and mapping similar terms with different specifications across the two organizations following the same general purpose algorithm proposed herein.

X. References

[1] N. F. Noy and M. A. Musen, "Algorithm and Tool for Automated Ontology Merging and Alignment," *Stanford Medical Informatics*, 2000.

[2] G. D. Simoni and R. Edjlali, "Magic Quadrant for Metadata Management Solutions," 8/15/2016 2016.

[3] E. Millard. (2010, Fruitful combination: A Successful Merger Requires Melding All Enterprise Data Within A Single Environment. *TeraData Magazine*. Available: http://www.teradatamagazine.com/v10n01/Features/Fruitful-combination/

[4] R. DeStefano, "Improving Enterprise Data Governance Through Ontology and Linked Data," DPS Dissertation, PACE, New York, 2016.

[5] C. Asamoah, "Knowledge Graph Syntax Validation and Visual Navigation for Developing Intelligent Systems," DPS Dissertation, PACE, New York, 2016.

[6] G. Alipui, "Reducing Complexity of Diagnostic Message Pattern Specification and Recognition with Semantic Techniques," DPS, Computer Science Pace, NY, 2016.

[7] I. William, O. N. Bonnie, and F. Lowell, *Business Metadata: Capturing Enterprise Knowledge*. New York, et. al.: Morgan Kaufmann Publisher, 2008.

[8] M. Uschold and M. Gruninger, "Ontologies and Semantics for Seamless Connectivity," *ACM SIGMOD Record*, vol. 33, pp. 58-64, 2004.

[9] J. Pollock. (2002, February 2002) Integration’ s Dirty Little Secret: It’ s a Matter of Semantics. *Modulant*.