Enterprise Knowledge Graphs: A Semantic Approach for Knowledge Management in the Next Generation of Enterprise Information Systems

Mikhail Galkin[1,2], Sören Auer[1], María-Esther Vidal[1,3] and Simon Scerri[1]
[1]*University of Bonn & Fraunhofer IAIS, Bonn, Germany*
[2]*ITMO University, Saint Petersburg, Russia*
[3]*Universidad Simón Bolívar, Caracas, Venezuela*

Keywords: Enterprise Information Systems, Linked Enterprise Data, Enterprise Knowledge Graphs, Semantic Web Technologies.

Abstract: In enterprises, Semantic Web technologies have recently received increasing attention from both the research and industrial side. The concept of Linked Enterprise Data (LED) describes a framework to incorporate benefits of Semantic Web technologies into enterprise IT environments. However, LED still remains an abstract idea lacking a point of origin, i.e., station zero from which it comes to existence. We devise Enterprise Knowledge Graphs (EKGs) as a formal model to represent and manage corporate information at a semantic level. EKGs are presented and formally defined, as well as positioned in Enterprise Information Systems (EISs) architectures. Furthermore, according to the main features of EKGs, existing EISs are analyzed and compared using a new unified assessment framework. We conduct an evaluation study, where cluster analysis allows for identifying and visualizing groups of EISs that share the same EKG features. More importantly, we put our observed results in perspective and provide evidences that existing approaches do not implement all the EKG features, being therefore, a challenge the development of these features in the next generation of EISs.

1 INTRODUCTION

The demand for new Knowledge Management (KM) technologies in enterprises is growing in recent years (Hislop, 2013). The importance of KM increases with the volumes of data processed by a company. Although the enterprise domain might vary from car manufacturing to software engineering, KM is capable of reducing costs, increasing the performance and supporting an additional added value to company's products. Novel KM approaches often suggest new data organization architectures and foster their implementation in enterprises. One of such an architecture leverages semantic technologies, i.e., the technologies the Semantic Web is based on, in order to allow machines to understand the meaning of the data they work with. Machine understanding and machine-readability are supported by complex formalisms such as description logics and ontologies. Semantic applications in the business domain comprise a new research trend, namely Linked Enterprise Data (LED). However, in order to truly exploit LED an organization needs to establish a knowledge hub as well as a crystallization and linking point (Miao et al., 2015). Google, for example, acquired Freebase and evolved it into its own Enterprise Knowledge Base (Nickel et al., 2016), whereas DBpedia assumed a similar position for the Web of Linked Data overall (Bizer et al., 2009).

In this paper, we present Enterprise Knowledge Graphs (EKGs), as formal models for the embodiment of LED. An EKG refers to a semantic network of concepts, properties, individuals, and links representing and referencing foundational and domain knowledge relevant for an enterprise. EKGs offer a new data integration paradigm that combines large-scale data processing with robust semantic technologies making first steps towards the next generation of Enterprise Information Systems (Romero and Vernadat, 2016). The main research goal of the paper is to formally define an EKG and provide an extensive study of existing Enterprise Information Systems (EISs), which offer certain EKG functions or can be used as a basis for an EKG. To achieve such a goal, we develop an independent assessment framework which is presented in the paper as well. We report on an unsupervised evaluation that allows for clustering existing EISs in terms of EKG features. Observed re-

Galkin, M., Auer, S., Vidal, M-E. and Scerri, S.
Enterprise Knowledge Graphs: A Semantic Approach for Knowledge Management in the Next Generation of Enterprise Information Systems.
DOI: 10.5220/0006325200880098
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 2*, pages 88-98
ISBN: 978-989-758-248-6

Table 1: Comparison of various Enterprise Data Integration Paradigms: P1:XML Schema Integration, P2:Data Warehouses, P3:Data Lakes, P4:MDM, P5: PIM/PCS, P6:Enterprise Search, P7:EKG. Checkmark and cross denote existence and absence of a particular feature, respectively.

Para-digm	Data Model	Integr. Strategy	Conceptual/ operational	Heterogen-eous data	Internal/ ext. data	No. of sources	Type of integr.	Domain coverage	Semantic repres.
P1	DOM trees	LAV	operational	✓	✓	medium	both	medium	high
P2	relational	GAV	operational	✘	partially	medium	physical	small	medium
P3	various	LAV	operational	✓	✓	large	physical	high	medium
P4	UML	GAV	conceptual	✘	✘	small	physical	small	medium
P5	trees	GAV	operational	partially	partially	✘	physical	medium	medium
P6	document	✘	operational	✓	partially	large	virtual	high	low
P7	RDF	LAV	both	✓	✓	medium	both	high	very high

sults give evidences that none of state-of-the-art approaches fully supports EKGs, and further study is required in order to catch the wave of future EISs.

The remainder of the paper is structured as follows: Section 2 positions EKGs into the ecosystem of EISs. Section 3 lays theoretical foundations of EKGs and formally describes the concept of EKGs. Section 4 presents the assessment framework and review of EISs that implement to a certain extent the EKG functionality. Further, the methodology followed to conduct the comparison is described, as well as an overview of current EISs and the description of the features necessary for implementing EKGs. Section 5 visualizes observed results using clustering algorithms and identifies hidden insights. Section 6 discusses data integration efforts in both the technical and enterprise dimensions, which might be considered as predecessors of EKGs. Section 7 analyzes the observed results and outlooks our future work.

2 MOTIVATION

In the last decades a variety of different approaches for enterprise data integration have been developed and deployed. Table 1 shows a comparison of the main representatives according to the following criteria: 1) *Data Model*: Various data models are used by data integration paradigms. 2) *Integration Strategy*: The two prevalent integration strategies are *Global-As-View* (GAV), where local sources are viewed in the light of a global schema and *Local-As-View* (LAV), where original sources are mapped to the mediated/global schema. 3) *Conceptual versus operational*: Some approaches primarily target the conceptual or modeling level, while the majority of the approaches are aiming at supporting operational integration, e.g., by data transformation or query translation. 4) *Heterogeneous data*: Describes the extent to which heterogeneous data in terms of data model or structure is provided. 5) *Internal/external data*: Describes whether the integration of external data is supported in addition to internal data. 6) *Number of sources*: Presents the typically provided number of sources, i.e., small (less than 10), medium (less than 100), large (more than 100). 7) *Type of integration*: States whether data is physically integrated and materialized in the integrated form or virtually integrated by executing queries over the original data. 8) *Domain coverage*: Reports the typical coverage of different domains. 9) *Semantic representation*: Expresses the extent of semantic representations, which can be low in the case of unstructured or semi-structured documents, medium in the case of relational or taxonomic data and high, when comprehensive semantic formalisms and logical axioms are supported.

The comparison according to these criteria shows, that the EKG paradigm differs from previous integration paradigms. EKGs are based on the statement-centric RDF data model, which can mediate between different other data models, since, for example, relational, taxonomic or tree data can be easily represented in RDF. Similar to XML-based integration and the recently emerging Data Lake approach, EKGs follow the Local-As-View integration strategy and are thus more suited for the integration of a larger number of possibly evolving sources. EKGs are the only approach, which bridges between conceptual and operational data integration, because on the one hand ontologies and vocabularies are used to model domain knowledge, but at the same time operational aspects, such as querying and mapping/transformation are supported. By employing RDF with URI/IRI identifiers, EKGs support the integration of heterogeneous data from internal and external sources either in a virtual or physical/materialized way. EKGs support the whole spectrum of semantic representations from simple taxonomies to complex ontologies and logical axioms. However, a promising strategy seems to be to use EKGs in combination with other approaches. For example, EKGs can be used in conjunction with a data lake, to add rich semantics to the low level source data representations stored in the data lake. Similarly, EKGs can provide background knowledge for enrich-

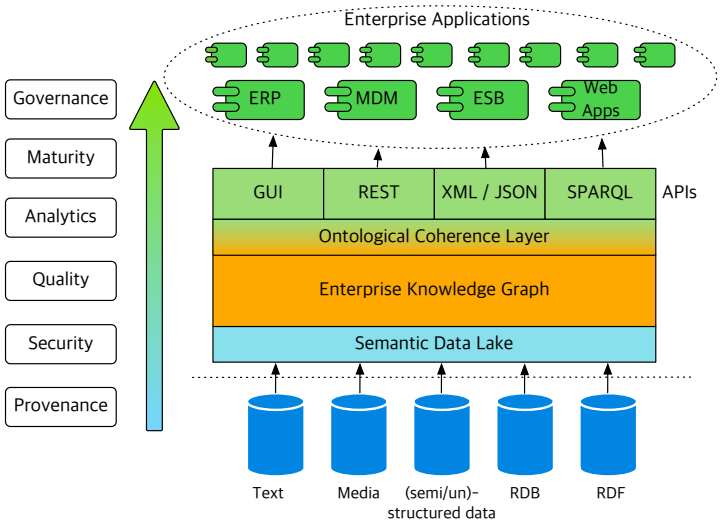


Figure 1: Position of an EKG in the Enterprise Information System architecture. An EKG assumes a mediate position between raw enterprise data storage and numerous services whereas the coherence layer provides a unified view on the data.

ing enterprise search or EKGs can comprise vocabularies and mappings from MDM.

Thus, EKGs occupy a unique niche in the ecosystem of enterprise applications. In order to provide a better understanding of such a niche, we present a structural view that represents a static orchestration of information systems in a company (cf. Figure 1).

The structural view shows an EKG as a consumer of raw heterogeneous data and a supplier of knowledge for enterprise applications. Numerous structured and unstructured data sources compose a new form of a data lake, i.e., a Semantic Data Lake (SDL). An SDL can integrate both company's private data as well as remote data in the open access, e.g., Linking Open Data Cloud. The Ontological Coherence layer contains a set of ontologies, both high-level and domain specific, which offer different semantic views on the EKG contents. For example, an organizational ontology defines a business department in a company, a security ontology places access restrictions on the data available to this department, whereas some supply chains ontology specifies the role of the department in a product lifecycle. Such a granularity increases knowledge representation flexibility and ensures that external applications use a standardized, ontology-based view on the required data from the EKG. The API layer exposes communication interfaces, e.g., graphical user interfaces (GUI), REST, XML, JSON, or SPARQL, to deliver knowledge encoded in EKGs to enterprise applications. Possible beneficiaries of using EKG technology include: Enterprise Resource Planning (ERP) systems, Master Data Management (MDM) systems, Enterprise Service Buses (ESB), E-Commerce and a company's Web applications. An EKG specifies a set of domain-independent features, e.g., provenance, governance, or security, which support the operation and maintenance of the entire system. We argue that these enterprise characteristics distinguish an Enterprise Knowledge Graph from a common Knowledge Graph; these features are explained in detail in the Section of EKG Technologies. Being 'orthogonal' to the knowledge acquisition and representation pipelines, the features are an integral part of the knowledge management policies. We also elaborate on those enterprise features in the Section of EKG Technologies.

3 ENTERPRISE KNOWLEDGE GRAPHS

We define Enterprise Knowledge Graphs (EKGs) as a part of an Enterprise Data Integration System (EDIS). Formally, EDIS is defined as a tuple:

$$EDIS = \langle EKG, S, M \rangle = \langle \langle \sigma, A, R \rangle, \langle OS, PS, CS \rangle, M \rangle \quad (1)$$

An EKG contains ontologies and instance data, S represents data sources, and M represents the collection of mappings to translate S into anEKG. $EKGs$ are based on the directed graph model $\langle N, E \rangle$ where nodes N are entities and edges E are relations between the nodes. An EKG is defined as a tuple $\langle \sigma, A, R \rangle$, where σ is a *signature* for a logical language, A is a collection of *axioms* describing an ontology, and R is a set of *restrictions* on top of the ontology.

A signature σ is a set of relational and constant symbols which can be used to express logical formu-

las. In other words, a signature contains definitions of entities, e.g., the RDF triples `:ToolX a owl:Class` and `:Product a owl:Class`[1].

The axioms in the set A provide additional description of the ontology by defining the relationships between the concepts defined in σ. In more detail, axioms leverage logical capabilities of the chosen ontology development methodology, e.g., RDFS allows for (but is not limited to) class hierarchy as well as domain and range definitions of properties. For instance, `:ToolX rdfs:subClassOf :Product`.

The restrictions R impose constraints on concepts and relationships. Although restrictions are also axioms, i.e., $R \sqsubset A$, we distinguish them as a separate important criteria necessary for large-scale multi-user enterprise environments. Restrictions can be expressed as triples or in a rule language, e.g., *Semantic Web Rule Language* (SWRL) or *SPIN*. Restrictions can be imposed on numerous characteristics: access rights, privacy, or provenance. To illustrate, suppose an RDF triple `:ToolX :cost 100` represents the cost of a tool `ToolX`. Then an RDF triple `:cost :editableBy :FinancialDept` is the restriction which states that only the financial department can change the value of the property `:cost`.

Data sources S are defined as a tuple $\langle OS, PS, CS \rangle$, where OS is a set of open data sources which an enterprise considers appropriate to re-use or publish, e.g., Linked Open Data Cloud or annual financial reports. PS is a set of private data sources of limited access. Supply chain data is an ample example of a private data. CS is a set of closed data sources with the strongest access limitations. Closed data is often available only to a special group of people within a company and hardly ever shared, e.g., technical innovations, business plans, or financial indicators.

Mappings in M connect data sources S with the EKG. The sources expose a semantic description of their contents. For instance, let $S_1 = \{producedFrom(x,y)\}$ return tuples that a certain product x is produced from a certain material y. Suppose EKG contains the following RDF triples: `{componentOf a rdf:Property. material a Class. product a Class}` . In order to query the tuples, one should provide mappings between the global ontology in the EKG and the sources. We advocate that the Local-As-View (LAV) paradigm is preferred in the $EDIS$. According to the LAV paradigm, sources are defined in terms of the global ontology (Ullman, 1997): for each source S_i, there is a mapping that describes S_i as a *conjunctive query* on the concepts in the global ontology EKG that also distinguish input and output attributes of the source. Therefore, M will contain a rule: $producedFrom(x,y) :- componentOf(y,x), material(y), product(x)$.
Although there exists a Global-As-View (GAV) paradigm which implies presentation of the global ontology concepts in terms of the sources, LAV approach is designed to be used with numerous changing sources and a stable global ontology. This is exactly the $EDIS$ case where an EKG remains stable but enterprise data is highly dynamic; thus, we underline the efficiency of LAV mappings for Enterprise Knowledge Graphs.

[1]We use "a" as shortcut for `rdf:type` as in the RDF Turtle notation for readability.

4 EKG TECHNOLOGIES

4.1 An Assessment Framework for Comparison

We propose an independent assessment framework to compare features and shortcomings of each surveyed technology. The benchmark categorizes high-level EKG functionality along three dimensions: D1) Human Interaction; D2) Machine Interaction; and D3) Strategical Development. Each dimension consists of a number of features. All the features of each dimension and their possible values are presented in Table 2. These features are later utilized in the review of EISs.

Human Interaction (HI) enables humans to interact with an EKG. Modeling expresivity characterizes the degree to support a user in modeling knowledge or a domain of discourse behind an EKG ranging from taxonomies and simple vocabularies to complex ontologies, axioms, and rules. Curation describes the support for creating, updating, and deleting knowledge from an EKG, and the availability of comprehensive user interfaces to perform such operations. Linking comprises the level of support for establishing coherence between knowledge structures and instance data. Exploration & Visualization functions are essential for a successful user experience. Among lay users (here referring to non-specialists in semantic technologies), an ability to explore data easily and represent it in a desired way often determines a solution's usability and success. Enterprise Search can be extended to allow for a semantic search over the entire enterprise information space, and beyond.

The Machine Interaction (MI) dimension describes different levels of support for machine interaction with an EKG. As information systems of the next generation, EKGs accentuate machine-to-machine in-

Table 2: Benchmark Dimensions (D) – D1: Human Interaction; D2: Machine Interaction; and D3: Strategical Development. Benchmark Parameters (P) – P1: Modeling; P2: Curation; P3: Linking; P4: Exploration; P5: Search; P6: Data Model; P7: APIs; P8: Governance; P9: Security; P10: Quality & Maturity; P11: Provenance; and P12:Analytics. Dash – no feature.

Dim.	#	Parameter	1	2	3	4	5	6
*D*1	*P*1	*M*odeling	Taxonomy	Thesaurus	Meta-schema	(SKOS) Vocabularies	Ontologies	Rules
	*P*2	*C*uration	Collaborative	plain text	forms	GUI	Excel	–
	*P*3	*L*inking	Text Mining	LOD Datasets	Ontology Mapping	Manual Mapping	Spreadsheet Linking	Taxo-nomy
	*P*4	*E*xploration/ Visualization	Charts	Maps	Web GUI	Faceted browser	Vis Widgets	Graphs
	*P*5	*S*earch	Semantic NL	Faceted	Federated Faceted	Full text semantic	Semantic search	–
*D*2	*P*6	*D*ata Model	RDF	RDF + docs	RDBMS	Property Graph	Taxonomy	
	*P*7	*A*PIs	SPARQL	REST	SQL	custom APIs	ESB & OSGi	Java/ Python
*D*3	*P*8	*G*overnance	Policies	Best practices	SAP	RDF based	–	–
	*P*9	*S*ecurity	Spring	ACL	SAP	Roles	Token	–
	*P*10	*Q*uality & Maturity	SKOS based	RDF based	SAP	–	–	–
	*P*11	*P*rovenance	Context	Data Lake based	SAP	–	–	–
	*P*12	*A*nalytics	Statistical	Exploratory	SAP	NLP	Big Data	–

Table 3: Benchmark-based Comparison of EKG Solutions. Parameter values are taken from Table 2: P1: Modeling; P2: Curation; P3: Linking; P4: Exploration; P5: Search; P6: Data Model; P7: APIs; P8: Governance; P9: Security; P10: Quality & Maturity; P11: Provenance; and P12:Analytics. Dash – no feature.

System	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*	*P8*	*P9*	*P10*	*P11*	*P12*
Ontorion	5,6	1,2,3	3	1,2	1	1	1	–	–	–	–	1,4
PoolParty	1,2,5	4	1,4	3,5	2	1	1,2,3	–	1	1	–	1,2
KnowledgeStore	4	–	1,2	4	1	2	1,2	–	–	–	1	–
Metaphacts	5,6	1,2,3	1,2	4,5	1	1	1	–	–	–	–	1
Semaphore 4	4	4	1,4	4,5	3	1	1,2,4	1	–	–	–	2,4,5
Anzo SDP	5	2,4,5	1,5	3,5	4	3	1,4,5	2	2	–	–	2
RAVN ACE	3	–	1	3,5,6	4	4	2,4,6		2,5	–	–	2,4
CloudSpace	5	–	2	4,6	1	1	1	–	–	–	–	2
ETMS	1,4	3,4	6	6	–	5	3	1	4	–	–	–
SAP HANA	–	–	–	–	–	4	3	3	3	3	3	3
EKG	5,6	1,4	1,2,3	3,5	4	1	1,2,4,5,6	4	2,4,5	2	2	1,2,5

teraction by creating a shared information space that is understandable by automated agents (e.g., services, software, other information systems). The Data Models feature specifies the foundational model of an EKG, i.e., how the knowledge is stored and represented in memory. We distinguish between Document models, Relational DBs, Taxonomies (Delphi, 2004), and the RDF data model. Whereas Data Models define the 'depth' of machine communication, available APIs expose the 'breadth' of the communication. An EKG is by definition involved in and connected with other information systems operating within an enterprise. APIs are a cornerstone and outline the difference between EKGs and previous generation of knowledge management solutions. The Strategical Development (SD) dimension is orthogonal to the HI and MI dimensions. Strategical Development features are of equal importance for both dimensions. Governance indicates availability of management mechanisms within a particular EKG implementation. EKGs for large enterprises should obey corporate policies and should be embedded into the decision-making routine. The Security feature is of high importance in large enterprises. EKGs should employ complex and comprehensive access control procedures, manage rights, and permissions on a large scale. The Quality & Maturity (Q&M) feature enables corporate management to evaluate the quality of an EKG and track its evolution over time. Data Quality indicates a degree of compliance of EKG data to an accepted enterprise standard level. Maturity shows a degree of applicability of a certain technology in an

enterprise. The Provenance feature allows for tracking origins from which the data has been extracted. Being applied to EKGs provenance elaborates on versioning and history of the EKG content. The Analytics feature provides statistics and overview of basic KPIs of an EKG. The variety of analytical services indicates aptitude of EKGs to be involved into enterprise information flows instead of being reluctant 'knowledge silos'.

4.2 EKG Solutions

Below, we compare 10 of the most prominent EKG available solutions. These are selected[2] based on their relevance given our own EKG definition, as well as their maturity (in a deployment-ready state and targeted for enterprise). The EKGs are evaluated against the benchmark presented in 4.1. Below we elaborate on the results summarised in Table 3.

Cognitum Ontorion[3] is a scalable distributed knowledge management system with rich controlled natural language (CNL) opportunities (Wroblewska et al., 2013). In the HI dimension, it enables sophisticated semantic modeling of high expressivity employing full support of OWL2 ontologies, semantic web rules, description logics (DL) and a reasoning machine. Collaborative ontology curation is available via a plain text editor and various forms. The platform supports ontology mapping out of the box although the type of mapping is not specified. Visualization and exploration means are presented by web forms and tables. CNL and semantic technologies stack enable natural language semantic search queries. In the MI dimension Ontorion employs RDF and SWRL to maintain the data model. Through the OWL API and SPARQL, machines can interact with the system. In the SD dimension Ontorion offers statistical analytics and Natural Language Processing (NLP) techniques.

Semantic Web Company PoolParty[4] is a software suite envisioned to enrich corporate data with semantic metadata in the form of a corporate thesaurus (Mezaour et al., 2014; Schandl and Blumauer, 2010). Knowledge can be modeled as a taxonomy, thesaurus, or as an ontology reusing built-in vocabularies, e.g., SKOS, schema.org, or custom. Curation is organized in the GUI via forms. Linking capabilities allow for text mining of unstructured data and manual vocabulary mapping. PoolParty provides web-based visualizations and exploration interfaces while traversing a taxonomy or ontology (visual browsing). SPARQL graphical shell is available to query the EKG. Semantic search supports facets, multilinguality, and structured and unstructured data including Sharepoint and Web CMS. To support MI, PoolParty uses the RDF model and implements JSON RESTful and SPARQL endpoints. In terms of SD features, Security relies on the Spring security mechanisms applied to the REST component of an EKG. Quality management is limited to SKOS vocabularies and checks potential errors and misuses. Analytical features perform statistical and exploratory data analysis.

KnowledgeStore[5] is a scalable platform (Rospocher et al., 2016) optimized for storing structured and unstructured data with the help of a predesigned ontology. KnowledgeStore is an integral part of the NewsReader[6] project. Modeling expressivity is limited to the predefined vocabulary so that newly acquired data must be represented according to this vocabulary. Linking functionality allows for automatic text mining and entity extraction with the subsequent juxtaposing the extracted data on the structured semantic data. A SPARQL client and a faceted browser are responsible for knowledge exploration. KnowledgeStore might be integrated with Synerscope Marcato[7] for comprehensive multi-dimensional visualizations. Semantic search over a graph might be performed using natural language queries or entities URIs. The data model employs RDF in conjunction with unstructured documents. One can interact with KnowledgeStore via REST API or SPARQL endpoint. Data sources might be tracked on the instance level, a limited form of versioning is implemented by the triples context feature.

Metaphacts[8] is a platform to design, control, and visualize domain-specific KGs. As to the HI, Metaphacts supports a number of features. Modeling opportunities rely on OWL ontologies and reasoning mechanisms. Collaborative knowledge curation is available in the Web-based GUI via forms and plain text editor. Interlinking capabilities allow for semantic enrichment of data extracted from unstructured documents. External public datasets might be used for enrichment similarly to KnowledgeStore. Metaphacts offers visualization and exploration in the form of a faceted browser with custom visualization widgets. Semantic search queries are executed against structured RDF data in the EKG. For the MI dimension Metaphacts reuses open standards RDF and OWL having SPARQL endpoint as a communication interface. However, a distinctive feature of the plat-

[2]The KMWorld Magazine served as a major source for their identification. http://www.kmworld.com

[3]Web: http://www.cognitum.eu/semantics/Ontorion/

[4]Web: https://www.poolparty.biz/

[5]Web: https://knowledgestore.fbk.eu/

[6]Web: http://www.newsreader-project.eu/

[7]Web: http://www.synerscope.com/marcato

[8]Web: http://metaphacts.com/

form is in the analytical domain as it provides machine learning and graph analysis functions enhanced by the GUI and custom visualization engine.

Smartlogic Semaphore 4[9] is a content intelligence platform which provides a broad spectrum of functions for an enterprise. The modeling is performed using taxonomies, SKOS vocabularies, and ontologies. Web GUI Editor allows for effortless curation leveraging Web forms, and drag & drop functions. Vocabularies might be interlinked on the schema level manually by a user. Data from unstructured documents is extracted and aligned with the vocabularies. Visualization and exploration means provide eloquent insights on the EKG and its construction routine. Text mining GUI guides a user from uploading a document to facts extraction. Semaphore 4 enables semantic search capable of answering federated queries and faceted queries. Elaborating on the MI Semaphore is oriented toward RDF data model. One of the important advantages for an enterprise is a set of integration interfaces with other enterprise-level solutions, e.g., Apache Solr, Oracle WebCenter Content. In the SD aspect, Semaphore allows for the creation of governance policies and user workflows to follow a certain policy. Powered by Big Data and NLP Semaphore offers a comprehensive analytical framework with rich visualizations.

Cambridge Semantics Anzo Smart Data Platform[10] is a data integration platform which allows structured and unstructured data to be applied in the corporate information flows. As regards HI, Anzo SDP's components allow for data curation, interlinking, rich visualizations, and advanced semantic search functionality. Modeling expressivity relies on fully-supported RDF and OWL ontologies. Curation is performed in a standalone Ontology Editor which exports the created ontology to the Smart Data Lake. SDP provides semi-automatic means for integration of Excel spreadsheets with ontologies with user involvement. SDP offers a comprehensive web-based GUI with a wide range of visualization widgets. Semantic search is supported by the NLP engine and is capable of operating on top of structured and unstructured content simultaneously. In the scope of MI Anzo SDP implements the RDF model built on top of Apache Hadoop HDFS, Apache Spark, and ElasticSearch. The platform might be integrated with existing IT solutions via numerous supported interfaces, i.e., Enterprise Service Bus (ESB), OSGi, SPARQL, JDBC, and custom APIs. As SD, the platform includes Governance, Security, and Analytics features. Governance is supported via default methodologies and best practices intended to preserve a robust workflow on each hierarchy level. A security mechanism is based on access control lists (ACL) and roles separation. Data analytics opportunities offered by SDP include exploratory analytics via interactive dashboards, sentiment analysis of text data, and spreadsheet analytics.

RAVN Applied Cognitive Engine[11] is an enterprise software suite for complex processing of unstructured data, e.g., text documents. ACE consists of numerous modules responsible for data extraction, search, visualization, and analytics. Being mostly a text mining tool, ACE modeling expressiveness adopts characteristics of the inner abstract schema. Yet the details of the schema remain unclear, the schema is used to store knowledge in the built-in graph database. Linking opportunities are presented with a broad spectrum of text analysis functions available in ACE. ACE provides a comprehensive web-based GUI to maintain the extraction and processing timeline. GUI presents interactive search, KG visualizations, provides widgets for the analytics module. Full-text semantic search in natural language is supported by Apache Solr and graph database. To support MI, ACE employs the property graph model on top of MongoDB and ApacheSolr[12]. The platform provides a range of REST, Java, and Python APIs to interact with enterprise applications. To support SD, ACE benefits from the comprehensive security subsystem and analytical functions. Security is based on token strings and access level. If a user token is contained in the allowed list, then her access level is determined by subparts of the token string. ACE analytical functions employ NLP to supply a user with sentiment analysis, expert locator, and exploratory visualizations.

SindiceTech CloudSpace Platform[13] provides a set of tools to build and maintain custom EKGs. CSP supports OWL ontologies for data modeling with high expressivity. The platform does not provide visual tools for curation. Linking functions allow for ontology-based integration when heterogeneous data from different sources is transformed to RDF and mapped with one ontology. Pivot Browser enables relational faceted browsing for data exploration. Ontologies can be visualized as graphs. Search function-

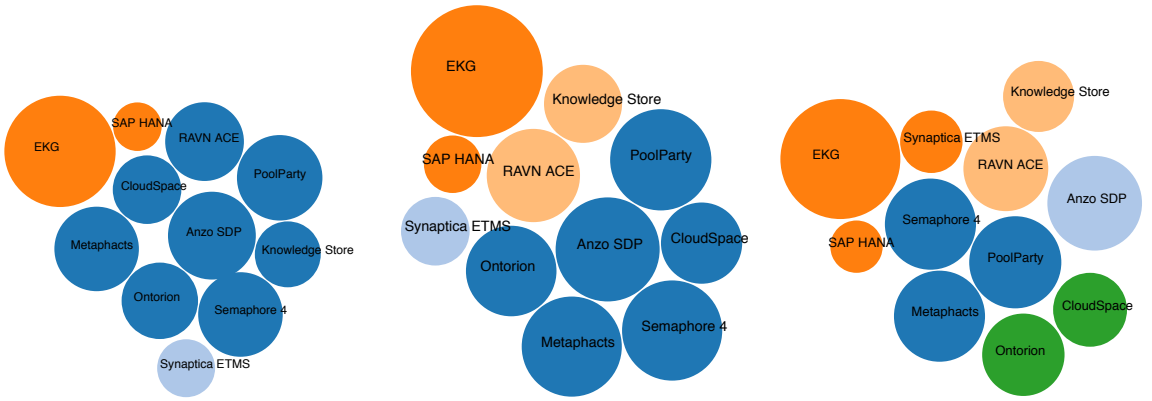
[9] Web: http://www.smartlogic.com/what-we-do/products-overview/semaphore-4

[10] Web: http://www.cambridgesemantics.com/technology/anzo-smart-data-platform

[11] Web: https://www.ravn.co.uk/technology/applied-cognitive-engine/

[12] Hoeke, J. V. Ravn cor whitepaper: https://www.ravn.co.uk/wp-content/uploads/2014/07/CORE-Whitepaper-W.V1.pdf

[13] Web: http://www.sindicetech.com/cloudspace-platform.html



(a) 3-clusters (b) 4-clusters (c) 5-clusters

Figure 2: (a) Hierarchical clustering, n=3. The orange cluster offers rich SD functionality, the azure corresponds to the taxonomy management, the blue comprises RDF based systems; (b) EM algorithm, n=4. All of the above, but the nankeen cluster emphasizes text mining; (c) K-means algorithm, n=5. The green cluster employs link discovery tools.

ality is powered by Semantic Information Retrieval Engine (SIREn) based on Apache Solr and ElasticSearch. In the MI dimension, CSP utilizes the RDF model with SPARQL as a basic API. SD features of the platform provide exploratory analytical services, e.g., modeling debugging and content overview.

Synaptica Enterprise Taxonomy Management Software[14] is a technology to manage corporate taxonomies, classifications and ontologies. ETMS modeling expressivity is limited to taxonomies and SKOS vocabularies. Curation means are based on a Web GUI which presents a tree-like interface with forms to fill in. Linking functionality is supported by taxonomy mappings. Mappings between taxonomies might be established automatically whereas mappings of vocabularies can be done manually via forms, and drag & drop in the GUI. Taxonomies and vocabularies are visualized as trees, hyperbolic graphs, hierarchies, area charts, and tree maps. For MI support, ETMS implements the Taxonomy model. The platform exposes RDBMS APIs to connect the system to the rest of the corporate IT environment. In the SD dimension, ETMS allows for data governance and security. Default policies and workflows specify 12 possible roles with certain access permissions. A user is categorized in one of such groups and has access to the constrained part of the taxonomy with limited rights.

SAP HANA SPS 11[15] is a "one size fits all" enterprise software suite which includes an opportunity to create and maintain EKGs on top of the HANA infrastructure and a specific graph database. HANA capabilities are overwhelmingly broad and cover all the features in the SD dimensions which are built-in in the default HANA distribution. However, an approach how HANA utilizes semantic technologies to support human interaction with an EKG remains vague. HANA Graph Engine is based on property graphs which naturally satisfy EKG needs. The engine exposes SQL-based APIs to communicate with versatile HANA modules and components.

EKG denotes the proposed EKG architecture with all the necessary features in HI, MI, and SD described in Section 3, Section 2, and Section 4.1.

5 EVALUATION OF STATE-OF-THE-ART

We conduct an unsupervised cluster analysis to existing EIS approaches for identifying groups of approaches of similar functionality[16]. Table 3 is transformed into a matrix with numerical values that denote normalized distances (in the range [0,1]) among the features, where values close to 1.0 indicate proximity to the defined EKG model. Values are chosen not to describe 'the best' or 'the worst' solution, but to stress the difference in functions the surveyed systems provide. *Weka*[17] is employed to perform the clustering into three, four, and five clusters. Varying the clusters number, we aim at identifying groups of EISs which share similar functions. One and two clusters give a superficial representation of the surveyed systems, and thus, are not included in the analysis. Three, four, and five clusters are able to reveal dependen-

[14]Web: http://www.synaptica.com/products/

[15]Web: https://hana.sap.com/capabilities/sps-releases/sps11.html

[16]Source codes are available in the github repository: https://github.com/migalkin/ekgs_clustering

[17]http://www.cs.waikato.ac.nz/ml/weka/

cies encoded in the vector representation which might have been overlooked or hidden during a manual review. Three clustering algorithms were applied, i.e., hierarchical algorithm, EM algorithm, K-means algorithm for three, four, and five clusters, respectively. We then visualized the results with *D3.js* visualization library. The radius of each node is proportional to the *Frobenius norm* of a features vector. The results are depicted in Fig. 2. The arrangement and positions of clusters are generated randomly, i.e., no cluster is logically closer to the nearest cluster than other clusters. Fig. 2(a) represents the distribution in three clusters computed by the *hierarchical* clustering algorithm. The biggest `blue` cluster comprises systems which can be described as RDF-based systems with rich visualizations, semantic search and analytical functions. However, the blue cluster lacks the features of the *Strategical development* dimension. On the other hand, the `orange` cluster contains systems that implement features from this dimension. The size of the envisioned *EKG* approach is bigger due to the leverage of the *Human Interaction* features which SAP HANA SPS lacks. The smallest `azure` cluster contains only one system which distinguishes itself as a taxonomy management tool.

Fig. 2(b) represents a distribution in four clusters performed by the *EM algorithm*. The EM (Expectation-maximization) algorithm involves latent variables to maximize likelihood of the input parameters. The extent of divergence in features (*SD* and taxonomies, respectively) keeps the `orange` and `azure` clusters the same as in the previous case. However, a new `nankeen yellow` cluster is marked off from the `blue` cluster. The newly discovered cluster is characterized by the emphasis on text mining functions of the *HI* dimension. The `blue` cluster is still defined as semantic-based systems with rich visualizations.

Fig. 2(c) represents the distribution in five clusters calculated by the *K-means algorithm*. The algorithm partitions input data in clusters by comparing means of an input vector and the clusters. The input vector is attached to the nearest mean value cluster. Firstly, the `blue` cluster was split further. The systems in the `blue` cluster are characterized as RDF-based knowledge management solutions which rely on ontologies, collaborative visual editing and wide range of supported APIs. Secondly, a new `green` cluster, derived from the `blue` cluster, is described as a set of systems that is based on semantic technologies and is capable of performing data interlinking. Additionally, the `green` cluster exposes only SPARQL and OWL API communication interfaces. The `nankin yellow` cluster remained the same as a distinctive set of text mining solutions. The `orange` cluster added one system. Nevertheless, the cluster is still described as an enterprise-friendly collection of systems with features from the *SD* dimension. The `azure` cluster now comprises a different system, Anzo SDP. The cluster contains a system capable of spreadsheet processing using ontologies for high-level schema definition. The wide range of supported APIs and RDB implementation of RDF distinguish the `azure` cluster.

6 RELATED WORK

6.1 The Technical Implementation Dimension

Data heterogeneity and volume are among the main challenges when building an EKG. Being comprised of numerous data sources, an EKG requires tools for interlinking, integration, and fusion of such data sources to ensure data consistency and veracity. Large volumes of common datasets (e.g., DBpedia dump is abput 250 GB, PubMed dump is about 1.6 TB) imply, that the integration pipeline has to be as automatic as possible. Prominent approaches for automatic linked data integration and fusion are LDIF (Schultz et al., 2012), OD CleanStore (Michelfeit et al., 2014), LIMES (Ngonga Ngomo and Auer, 2011). All of them maintain an integration pipeline starting from data ingestion from various remote data sources to a high-quality target data source. LDIF and ODCS resort to SILK (Isele and Bizer, 2013) the tasks of entity recognition and link discovery tool. LDIF employs Sieve (Mendes et al., 2012) as the Data Fusion module which aims at resolving property values conflicts by the assessment of the quality of the source data and by application of various fusion policies, while OD CleanStore uses a custom data fusion engine. LIMES (Ngonga Ngomo and Auer, 2011) is a stand-alone link discovery tool which utilizes mathematical features of metric spaces in order to compute similarities between instances. Sophisticated algorithms significantly reduce the number of necessary comparisons of property values. Linked Data integration tools digest data from various public RDF datasources. Those community driven knowledge graphs are either general-purpose bases (e.g., DBpedia, Wikidata[18], Freebase) which comprise facts from different domains or domain specific bases which aim at providing detailed insights on a particular theme. DBpedia is one of the largest graphs containing more than three billion facts. Wikidata is envisioned to be a universal source of information for populating

[18]https://www.wikidata.org/wiki/Wikidata:Main_Page

Wikipedia articles. Freebase has been acquired by Google as a source for its Knowledge Graph and Knowledge Vault (Dong et al., 2014).

6.2 The Enterprise Dimension

Large enterprises have integrated Semantic Web technologies with their IT infrastructures in various forms. Statoil used Ontology-Based Data Access (OBDA) to integrate their relational databases with ontologies (Kharlamov et al., 2015). NXP Semiconductors transformed product information into an RDF product taxonomy (Meenakshy and Walker, 2014). Stolz et al. (Stolz et al., 2014) derived OWL ontologies from product classification systems. All the abovementioned efforts have, however, only analyzed the one specific domain that is directly relevant, and although they refer to KGs profusely the authors do not elaborate on the definition, conceptual description or reference architecture of such KGs. In general, a precise definition of the EKG concept remains to be developed. One contribution of this paper is to address the above by clearly defining the concept, and positioning the potential of EKGs. EKGs have barely been explored along the data quality dimension in either Business Informatics or Semantic Web communities. The Competence Center Corporate Data Quality (CC CDQ) has developed a framework (Otto and Oesterle, 2015) for Corporate Data Quality Management (CDQM) that presents a model for the evaluation of data quality within a company. Being a thoroughly tailored enterprise tool, the framework is, however, only applicable to conventional data architectures, i.e., MDM, Product Information Management (PIM) and Product Classification Systems (PCS), and extending the framework to cover semantic data architectures is impractical.

7 CONCLUSIONS AND FUTURE WORK

In this article, we presented the concept of Enterprise Knowledge Graphs, described its structure, functions, and purposes. We devised an assessment framework to evaluate essential EKG properties along three dimensions related to the human interaction, machine interaction, and strategical development. We provided an extensive study of existing enterprise solutions, which implement EKG functionality to a certain extent. The analysis indicates a wide variety of approaches based on different data, governance, and distribution models and emphasizing the human and machine interaction domains differently. However, the strategical development functions, i.e., Provenance, Governance, Security, Quality, Maturity, which are of utmost importance for an enterprise, often still remain unaddressed by current EKG technologies. We see here a room for significant improvements and new innovative technologies empowering companies to fully leverage the EKG concept for data integration, analytics, and the establishment of data value chains with partners, customers and suppliers. In future work, we aim at enhancing the EKG model, implement a reference EKG architecture, and develop a prototype which implements and supports the next generation of EISs.

REFERENCES

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, 7(3):154–165.

Delphi (2004). Information intelligence: Content classification and the enterprise taxonomy practice. Whitepaper.

Dong, X., Gabrilovich, E., Heitz, G., and Horn, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *20th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*, pages 601–610.

Hislop, D. (2013). *Knowledge management in organizations: A critical introduction*. Oxford University Press.

Isele, R. and Bizer, C. (2013). Active learning of expressive linkage rules using genetic programming. *Web Semantics*, 23:2–15.

Kharlamov, E., Hovland, D., Jiménez-Ruiz, E., Lanti, D., Lie, H., Pinkel, C., Rezk, M., Skjæveland, M. G., Thorstensen, E., Xiao, G., et al. (2015). Ontology based access to exploration data at statoil. In *The Semantic Web-ISWC 2015*. Springer.

Meenakshy, P. and Walker, J. (2014). Applying semantic web technologies in product information management at nxp semiconductors. In *13th International Semantic Web Conference (ISWC 2014)*.

Mendes, P. N., Mühleisen, H., and Bizer, C. (2012). Sieve: Linked data quality assessment and fusion. In *2012 Joint EDBT/ICDT Workshops*, pages 116–123.

Mezaour, A.-D., Van Nuffelen, B., and Blaschke, C. (2014). Building enterprise ready applications using linked open data. In *Linked Open Data–Creating Knowledge Out of Interlinked Data*, pages 155–174. Springer.

Miao, Q., Meng, Y., and Zhang, B. (2015). Chinese enterprise knowledge graph construction based on linked data. In *Semantic Computing (ICSC), 2015 IEEE International Conference on*, pages 153–154. IEEE.

Michelfeit, J., Knap, T., and Nečaskỳ, M. (2014). Linked

data integration with conflicts. *arXiv preprint arXiv:1410.7990*.

Ngonga Ngomo, A.-C. and Auer, S. (2011). Limes - a time-efficient approach for large-scale link discovery on the web of data. In *IJCAI*.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Otto, B. and Oesterle, H. (2015). *Corporate Data Quality. Prerequisite for Successful Business Models*. epubli.

Romero, D. and Vernadat, F. B. (2016). Future perspectives on next generation enterprise information systems. *Computers in Industry*, 79:1–2.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., and Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Schandl, T. and Blumauer, A. (2010). Poolparty: Skos thesaurus management utilizing linked data. In *The Semantic Web: Research and Applications*. Springer.

Schultz, A., Matteini, A., Isele, R., Mendes, P. N., Bizer, C., and Becker, C. (2012). Ldif-a framework for large-scale linked data integration. In *21st Int. World Wide Web Conference (WWW 2012), Developers Track*.

Stolz, A., Rodriguez-Castro, B., Radinger, A., and Hepp, M. (2014). Pcs2owl: A generic approach for deriving web ontologies from product classification systems. In *The Semantic Web: Trends and Challenges*, pages 644–658. Springer.

Ullman, J. D. (1997). Information integration using logical views. In *International Conference on Database Theory*, pages 19–40. Springer.

Wroblewska, A., Kaplanski, P., Zarzycki, P., and Lugowska, I. (2013). Semantic rules representation in controlled natural language in fluenteditor. In *Human System Interaction (HSI), 2013 The 6th International Conference on*, pages 90–96. IEEE.