A Blueprint for Semantically Lifting Field Trial Data: Enabling Exploration using Knowledge Graphs

Burkow, D.[1#], Hollunder, J.[1#], Heinrich, J.[1], Abdallah, F.[1],
Rojas-Macias, M.A.[1], Wiljes C.[2], Cimiano P.[2,3], Senger, P.

[1] Bayer CropScience, CLS Translational R&D, Germany
[2] Semalytix GmbH, Bielefeld, Germany
[3] Semantic Computing Group, Bielefeld University, Germany
[#] Both authors contributed equally to this work.
jens.hollunder@bayer.com, cimiano@semalytix.de

Abstract. In many organizations, relevant data is distributed across multiple primary data sources using different data models, schemas, and vocabularies which prevents effective knowledge management at an organizational level. As a result, it is very cumbersome to retrieve and integrate relevant data necessary to answer specific business questions. We propose a semantic lifting architecture that lifts primary data resources into a single corporate knowledge graph, implemented following FAIR and Linked Data principles, which assigns unique URIs to entities and maps the data using a uniform shared vocabulary. This structure supports integration of heterogeneous and distributed data sources to enable advanced search and analysis. In this paper we consider the case of field trial data collected and managed by Bayer, present an example of our architecture, and describe an application using this architecture that provides semantic search and visual exploration of field trial data to relevant stakeholders. As a main benefit, our architecture allows domain experts and analysts to retrieve timely and consistently labeled data originating from multiple primary sources while maintaining data context to enable cross-domain analytics. This provides the means to answer more complex questions with minimal data preparation, improving speed and precision by driving decisions with semantically lifted data.

Keywords: Semantic Data Modeling, Knowledge Graph, Linked Data, Data Science, Visual Analytics, Integrative Reasoning, Digitalization.

1 Introduction

In modern and agile organizations, a key challenge is to exploit data and related knowledge assets to support timely and informed decision making. Data analytic solutions, however, require consistent, integrated and semantically well-defined data. Such integrated data are often not available in large organizations due in large part to the fact that data are typically confined to siloed legacy storage systems where the primary data

CEUR-WS.org/Vol-2849/paper-08.pdf

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

was captured[1]. Among other things, conflicting database schemas, aggregation levels, and disparate terminologies hinder seamless data science solutions when working across data repositories.

At Bayer CropScience - as in many other similar organizations - multiple and heterogeneous datasets exist which are stored in raw data repositories to fit the needs of the different steps within the research and development process. This data ranges from early molecule research, greenhouse- and field-testing, all the way to regulatory, marketing and administrative data. Working with these core data assets has a huge potential for value generation using translational data science approaches: the potential to generate novel insights to improve current business models, streamline and optimize R&D processes, and support data-driven decision-making at a whole company level. Translational data science approaches help prevent duplications of effort, optimize investments with organizational learning from previous experiences, and drive better cost and risk evaluations.

In this paper, we focus on the case of lifting and organizing field trial data to support semantic exploration and discovery of those data. A common existing use case by product development managers includes searching and filtering by various trial properties and statistically analyzing the resulting data – a process we have streamlined and improved with semantically integrated data. The data are stored in different primary data sources that use different schemas, different coding conventions, and different aggregation levels. For example, a single chemical can be referred to by various names across data sources. This makes querying the data across data sources challenging and time consuming. In addition, there are misspellings and missing values in the data, and different column and field naming conventions are used across departments and groups, rendering the problem of seamless data exploitation every more challenging.

In this paper we propose a semantic lifting approach which we have implemented to harmonize and integrate field trial data from multiple primary data resources to support uniform access and querying using controlled vocabularies. For this, we propose a semantic lifting architecture that supports data cleaning and data normalization by mapping data to a corporate knowledge graph. As a net result, domain experts and product managers can answer queries over all the primary data sources in an integrated fashion. The semantic lifting approach further increases the interoperability of the data and supports a unified quality assessment strategy that allows to improve the quality of the data, thus increasing the reusability of the data, and supports the application of data mining and analytics technology on the integrated data.

Our semantic lifting approach relies on a semantic data layer that integrates the disparate resources into a company-wide knowledge graph for field trials. This approach integrates a variety of sources, uses a collection of domain-specific, semantic vocabularies to foster harmonization, and supports multiple output formats to enable unparalleled flexibility in data utilization. In section 2 we describe our approach to semantic data lifting, in section 3 we demonstrate the benefits of this semantic integration in facilitating semantic search and exploration of field trial data, we discuss further benefits of our approach in the conclusion, and finally, we point toward future work.

[1] https://www.elderresearch.com/blog/42-v-of-big data, last accessed 2019/02/20

2 A knowledge graph enabled approach to semantic data normalization and lifting

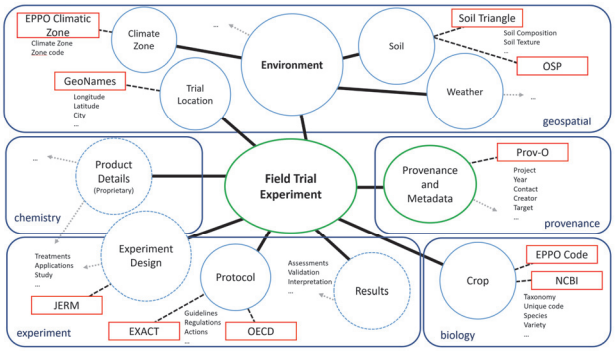

Fig. 1. Field trial data model snapshot based on the selected use case: Overview of pseudo-schema used to define semantic ontologies over the data, categorized into different domains (geospatial, chemistry, provenance, experiment, and biology). Red objects denote external, regulated ontological libraries and terminology (Table1 for more details). Dotted objects are category placeholders for connections to entities not relevant to this paper.

The main goal of our proposed semantic approach to knowledge organization is to support advanced data analytics facilitating better reusability, explainability and discoverability of contextualized business data. Generally, data analytic approaches often require the ability to consume a huge amount of data (e.g. in order to train learning models properly [1]). Hence, a common bottleneck is still the time needed for ad-hoc data preparation and curation, leading to the claim that 80% of a data scientist's time is invested in data integration, cleansing, and alignment – colloquially known as the 80/20 rule[2] – and surveys confirm that the actual time investment is at least 50%[3]. Fortunately, semantic data and knowledge representations can easily deliver the necessary quantity of data and can significantly reduce the amount of preparation time by integrating and cleaning the data once for many later applications. Linked Data offers an agile, flexible solution for data integration by using a schema-less graph-based data representation [2]. The general approach for semantic lifting consists in lifting data from individual data silos into a knowledge graph representation guided by a semantic data "domain model" and re-using identifiers from predefined shared vocabularies (example Fig.1). For this purpose, the data needs to be organized and governed over their lifecycle in such a way that different key quality aspects, such as completeness, validity, accuracy, granularity, and consistency are met[4]. Central to our proposal are four key elements: (a) a corporate knowledge graph representing all the knowledge available within the organization harmonized over (b) domain-specific vocabularies as keys to (c) polyglot knowledge management [3] while following Linked Data principles to support data

[2] https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html
[3] https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf
[4] https://www.blazent.com/seven-characteristics-define-quality-data/, last accessed 2019/09/10

FAIRness[5] [4] and generate (d) multiple views orchestrated by various knowledge representations. Those representations provide task-specific workspaces with cross-domain sets of lenses to view and to analyses data from different perspectives.

2.1 Approach to semantic lifting

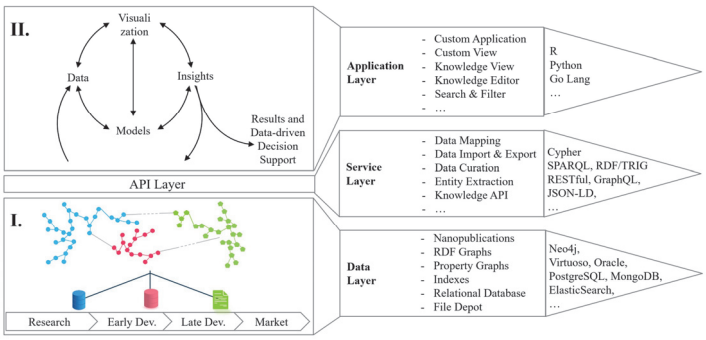

Fig. 2. Decision-making across the development pipeline: (I) **Semantic Lifting**: semantic data modeling across domains, augmenting internal and external data resources, ensuring data quality, and generating knowledge assets on top of different backend data solutions; (II) **Advanced Data Science:** data retrieval and exploration, visual analytics [6], machine learning and statistical models, all encapsulated within an iterative approach [7] resulting in data-driven decision making by utilizing the generated interpretations, implications, and conclusions.

Our approach lifts data assets from their primary sources into a semantic knowledge layer with contextual meaning and establishes connections to other related datasets. This makes the data easily searchable, analyzable, and reusable [4]. Data lifting relies on the domain model structure and internally agreed upon controlled vocabularies (Table1), which capture the semantic meaning of each datum and allow us to generate domain-specific knowledge representations that are seamlessly linked. Following the common standard applied by organizations such as Google[6] or by public data resources such as Wikidata[7], we implement knowledge graphs to construct the semantic knowledge layer. Knowledge graphs rely on a graph data structure to represent knowledge, whereby nodes are connected to each other by edges to capture the relations between data entities. We include properties and metadata from the original source data as well as external sources. Although there currently is no web standard for property graph representations, there are ongoing W3C activities[8] to close this gap and to push labeled property graphs towards standardization as well as other solution implementations bridging between those solutions [5].

[5] https://www.force11.org/group/fairgroup/fairprinciples, last accessed 2019/02/20
[6] https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html
[7] https://www.wikidata.org/, last accessed 2019/02/20
[8] https://www.w3.org/Data/events/ data-ws-2019/, last accessed 2019/02/20

Our semantic lifting pipeline is embedded in our end-to-end technology stack (Fig.2) and is composed of **producers, transformers**, and **consumers** within the corresponding data layer (Fig.3). The data layer contains a variety of distinct components according to requested user requirements and is connected to an application layer through a service layer. The service layer uses micro- and web-service technology to provide standardized exchange formats and functionality for integration, curation, interaction, and analytics. It also provides direct consumption by end user applications (such as in visual analytics tools) and serves data scientists. The primary focus of the paper is demonstrated in part I of Fig.2, whereas the provided use case is going one step further by accessing the generated knowledge representations through the API layer, demonstrating some of the aspects of the data science cycle of part II.

Table 1. Used ontologies and terminologies for the selected use case.

Context	Terminology	Source
Provenance and Metadata	Provenance Ontology (PROV-O)	W3.org
	Terms for field trial provenance description	*Internal*
Environment (general, soil, climate zone, geolocation, field, plot, etc.)	EPPO Climatic Zone	EPPO.int
	Ontology of Soil Properties and Processes (OSP)	*Du et al., ISWC 2016*
	GeoNames Terminology	GeoNames.org
	Soil Triangle Terminology	USDA.gov
	Terms describing trial sites, field, plots, etc.	*Internal*
Crop (species, variety, etc.)	NCBI Taxonomy of Species	NCBI.NLM.NIH.gov
	EPPO Code	EPPO.int
Experiment (experimental design, setup, protocol, results, etc.)	Plant Experimental Conditions Ontology (PECO)	BioPortal.BioOntology.org
	Crop Ontology (CO)	cropontology.org
	OECD guidelines & Harmonised Templates	OECD.org
	EXperimental ACTions (EXACT)	BioPortal.BioOntology.org
	JERM Ontology	jermontology.org
	Terms describing field trial experiments	*Internal*

Producers: The semantic workflow starts with the ingestion of relevant data to incorporate into the corporate knowledge graph. We have developed various source connectors, called producers, which allow us to communicate with each of the different data sources. Thus, we interact with the data from the original source, which remains fully disparate and federated, and we format it to allow for easy manipulation by downstream transformers. Additionally, we implement a feedback process to ingest results from prior analysis and related decisions through an additional producer to augment the primary data.

The data are cleaned by, for instance, formatting numeric and date fields to fit the defined needs in the domain model, some missing data is simply imputed by various methods, and some fields are merged to match the destination formatting requirements. Depending on the data source, this process can be done automatically with simple scripts, or manually (i.e. case by case) with the help of domain experts. Manual

corrections and revisions are stored as explicit rules to maintain consistency over database updates to ensure reusability.

The key is creating or maintaining data consistency and cohesion across data sources so we can interlink the data and enrich or augment it with other resources or services within our life science business domain. The producers become a *de facto* process for monitoring quality of source data and help to discover inconsistencies or other data issues that can be flagged or forwarded to the owner of the original data to be corrected.

Transformers: The second step in the workflow requires enforcing controlled vocabularies and a common semantic data model to integrate the previously heterogeneous datasets. The set of used vocabularies are composed of internally collected terminology lists as well as external ontologies which can be completely or only partly integrated depending on our needs (Table1).

Most incorrect or inconsistent data are detected through automatic plausibility and consistency checks during source data ingestion and is corrected across all the relevant data producers. The data are then combined and passed through a collection of semantic connectors to transform the source data into knowledge representation formats. The transformers are modular and reusable components that perform the translation, enrichment, and augmentation of internal and external data resources to create the corporate knowledge graph. Some transformers are source specific, through which data from only a subset of producers are sent, while other connectors are data type agnostic using technology such as R2RML[9] to map the relational data consistently to the graph structure.

The internal ontology platform is a key element for the final data model. It hosts ontologies for relevant domain knowledge, terminology concepts, and structured metadata that describe provenance, support, evidence, and workflow (e.g. ontologies for instance extracted from BioPortal[10]). It is the primary source of semantic information used during the lifting process. Maintaining such a reference database enables trust, versioning, and traceability management for the dynamic knowledge graph.

For an explicit example, a key transformer that many downstream transformers rely on is our “location service” transformer. This transformer checks, cleans and corrects all coordinate information to ensure they match the stated location. For instance, we infer an approximate location for historical datapoints with no GPS coordinates based on trial site metrices, and, at worst, the GPS coordinates from the closest verifiable city are used. This kind of data is a prerequisite to obtain correctly mapped and modeled climate, soil, or weather details from external sources in downstream transformers.

Consumers: The data is virtualized into a variety of target database structures after it has been lifted into the semantic knowledge layer by consumer modules. The most flexible target database, and the one most representative of the corporate knowledge graph, is a graph structure stored in Virtuoso and/or Neo4j. Additionally, we implement a MongoDB service to provide file-based persistence for some data assets, and ElasticSearch indices are produced to enable filtering and semantic (fuzzy) searching on content, provenance, and metadata (see section 3). Additionally, some business data

[9] https://www.w3.org/TR/r2rml, last accessed 2019/02/20
[10] https://bioportal.bioontology.org/, last accessed 2019/02/20

(e.g. project information) are virtualized within PostgreSQL for easy ingestion by existing applications.

A service layer encapsulates the knowledge base, connecting to the backend-solutions of the data layer to enable consumption via provided applications or APIs using GraphQL[11] or REST[12]. Thus, the consumers provide a suite of methods to access the knowledge base providing multiple formats for various applications. We use the GraphQL API and the ElasticSearch indices to feed search and visual analytics dashboards that management can directly interact with (e.g. Tableau, SpotFire, or scripts based on D3.js[13]). At the same time, using the same API, data scientists can refresh their analytical tools with the most up to date, semantically enriched data. And at any time, new consumers can be developed to handle novel requirements from new use cases without modifying the underlying corporate knowledge graph.

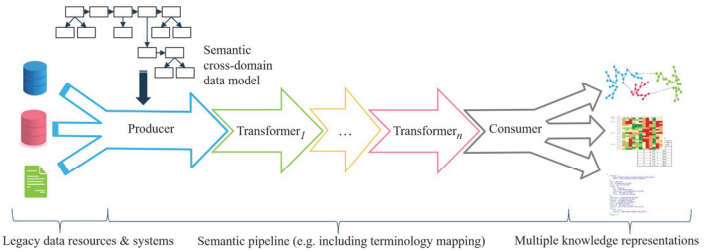

Fig. 3. General semantic lifting from identified data sources via several transforming steps to unified polyglot semantic knowledge representations.

3 Semantic search and exploration over enriched and integrated internal field trial data

The proposed architecture has been implemented within our agricultural domain, serving decision making applications used by different stakeholders, such as scientists, project managers, or decision makers. In the Field Solutions department, various products and equipment are tested on fields around the world for reasons spanning from dose-finding and efficacy compared to current standards to market development and regulatory validation. Additionally, the global nature of our field trial experiments introduces regional operational differences that need to be fixed before cross-trial comparisons can be made, differences such as subjective judgement scales or measurement units.

We developed an in-house question-answering application to search for and analyze field trial experiments. As previously described, the knowledge representation is generated by ingesting data though **producers** from a number of distinct repositories and resources, such as an internal field trial database (e.g. crop, product, location, assessment, observation, etc.), field sensor-data repository (e.g. Holland Scientific

[11] https://graphql.org/, last accessed 2019/08/01
[12] https://restfulapi.net, last accessed 2019/08/01
[13] https://d3js.org/, last accessed 2019/02/20

CropCircle[14] or various other sensors), and internal or external environmental data resources (e.g. soil, climate zones, weather, disease pressure, etc.). The **transformers** implement expert knowledge to fix the inconsistencies and sanity-check the results. The resulting field trial knowledge graph generated by the **consumer** can then be accessed using the GraphQL API for any number of downstream applications.

For reference, the Neo4j field trial subgraph is just a part of the entire knowledge graph, but it comprises 5.3 billion relations for 600 million nodes describing more than 500,000 field trial experiments, and the graph is continuously growing. The domain model is an integral key to enable iterative growth since the connection points between nodes belonging to individual experiments are explicitly defined, and unique nodes between experiments remain unique. This methodology differs from social network graphs where the structure can dramatically change from week to week, whereas our methodology provides a stable structure as time progresses and new information is incrementally added. Thus, this lifted data provides opportunities for visualization and designing predictive models which, together with graph-based machine learning approaches, improve our R&D analytics and trial planning pipelines.

In the past, product development managers used to interact with a single software access point to a cumbersome relational database to extract Excel reports that they manually formatted for results meetings. However, we developed a consumer to produce a robust, simple relational database that is used by Spotfire to compare trials within a series, across multiple trial series, and even over multiple years – all without any need for Excel exports or manual formatting. These visual analytics tools are the basis for many managers' results discussions slide decks, and there is almost no repeated effort to update these graphics as new trial information comes in from the field. What used to take up to one hour per trial to generate a human readable Excel product performance overview now takes less than 20 minutes for an entire trial series across multiple years.

We also embedded fuzzy search (using ElasticSearch) into a web-hosted field trial overview application to enable real-time, type-ahead functionality to search for and filter field trials based on any of the many trial properties (such as weather or assessment type) in our knowledge graph. The search API feeds a visualization interface developed in JavaScript and Angular to provide a custom, modern, web-based experience to the end-user with drill-down data selection based on controlled terminologies managed through our ontology platform. This means that a product manager can access the same information by searching for the marketed product name, the in-house mixture code, or the active ingredient name of a formula due to the synonym repository appended by one of the pipeline transformers during data lifting. This web-based architecture is versatile in that it allows for seamless integration of dashboards from third-party vendors such as Tableau or Spotfire in addition to custom, task-specific dashboards and visualizations – something that would have been almost impossible to achieve with the original database landscape. Each dashboard can be access-controlled to allow users with various requirements (e.g. scientists, projects managers, or even members of leadership teams) access to appropriate organizational data, all from a single online portal. Questions varying from "how many trials are currently planned in Europe" to "what are the

[14] https://hollandscientific.com/portfolio/crop-circle-phenom/, last accessed 2019/04/07

sensor measurements from trials using product A with foliar application in a field with a soil pH greater than 7" can be answered just as simply. In contrast, the second question would have required three separate data exports from two databases and some wrangling to merge and filter the results before the question could even be answered, let alone visualized. This kind of dashboarding portal reduces the effort necessary for results overview discussions, trial planning and resource allocation, and other business critical R&D operations. Overall, the semantic pipeline has given us a higher quality data standard, ensured comparability and analytical reproducibility [4], and improved precision and recall on search queries for our users. Other applications that we have developed using this pipeline blueprint include a Nanopublications [8, 9] subgraph with connections to key omics data and a chemical trials subgraph for small molecules experiments as well as knowledge graph refinement and analytic approaches [10, 11].

4 Discussion and conclusion

We described the architecture for knowledge organization used at Bayer Crop Science R&D as well as the status of a prototypical implementation of these principles. Our approach relies on a semantic data integration layer that exists on top of, and in conjunction with, the underlying data repositories existing across the organization. The primary data are lifted into a semantic corporate knowledge graph layer based on a prescribed domain model. This enables the organization to choose which subset of data dimensions to lift, thus following an agile approach that limits required overhead costs and reduces data redundancy as much as possible. New data sources can be attached to the already linked data without disturbing the existing framework, and data collection workflows in each separate division remain unaltered.

Semantic integration of data requires the adherence to vocabularies that are shared across stakeholders and organizational units to support data normalization. Adopting shared vocabularies between divisions increases conceptual consistency and reduces ambiguity, enabling analytical tools and results discussions to be shared between departments. A key benefit is the unifying effect of using consistent and persistent identifiers across primary data sources which eliminates specificities inherent to single data sources and represents data in a source-independent way, reducing redundancy without losing context. Such a semantics-driven data integration approach has demonstrated numerous benefits: connections between data from different sources are clarified, distinct data silos are aligned, and interdepartmental data are contextually comparable. Knowledge is represented in a context that reflects a consistent view across the organization, allowing abstractions and aggregations that were previously infeasible. Semantically lifted data are easily bundled into domain-specific data assets and exposed via APIs which enable existing workflows to continue working seamlessly on top of the lifted data landscape. Cross-functional data assets can be seamlessly accessed and used for machine learning, dashboarding, and visual analytics via use-case specific APIs in a time-saving way.

An important benefit of a semantic integration approach is that it contributes to more complete, consistent and bias-free insight generation from data. Some inconsistencies

or gaps, such as disconnects between the needs of later regulatory trials and early product phase trial planning, can be discovered early on in our data integration layer, at the knowledge graph level. Augmenting original data with new inferred data adds new depth and enforces the business context with derived data that is made persistent. Adding simple reasoning mechanisms at the data lifting stage eliminates the need to calculate such metrics regularly at the data query or analysis stages, reducing time to analysis and computational load. Additionally, for common calculations, this implementation ensures consistency between downstream analytical approaches across divisions.

As this methodology is continually refined, further development of the outlined approaches will focus on the evolution and modification of the data and shared vocabularies and on implementing a layer for quality control via semantic unit tests (e.g. using RDFUnit [12]) that involve the data owners. Additionally, we would like to expand on the topic of implementing a pay-as-you-go data lifecycle that allows project scaling and need-based investment increase. The next steps also include expanding and maturating the data model by challenging it with additional use cases, focusing on data provenance, analytics, machine learning on knowledge graphs, and advanced integrative reasoning.

5 References

1. Bhatnagar, R.: Machine Learning and Big Data Processing: A Technological Perspective and Review, DOI: 10.1007/978-3-319-74690-6_46 (2018).
2. Wiljes, C., & Cimiano, P.: Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology. Proceedings of the Workshop on the Semantic Publishing (2012), 48-59.
3. Dieter Fensel: Ontologies - a silver bullet for knowledge management and electronic commerce. Springer 2001, ISBN 978-3-540-41602-9, pp. I-IX, 1-138.
4. Wilkinson, M.D., et al.,: The FAIR Guiding Principles for scientific data management and stewardship, SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18.
5. Thakkar H., Punjani, D., Lehmann, J., Auer, S.: Two for one: querying property graph databases using SPARQL via gremlinator, GRADES-NDA'18, DOI: 10.1145/ 3210259.3210271.
6. Chung Wong, P., Thomas, J.: Visual Analytics. In IEEE Computer Graphics and Applications 24 (5), 20-21 (2004).
7. Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F.: Mastering the Information Age – Solving Problems with Visual Analytics. Eurographics Association (2010).
8. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use 30(1-2), 51–56 (2010) Gibson et al., 2012.
9. Kuhn T., Merono-Penuela, A., Malic, A., Poelen, J.H., Hurlbert, A.H., Centeno Ortiz, E., Furlong, L.I., Queralt-Rosinach, N., Chichester, C., Banda., J.M., et al.: Nanopublications: A Growing Resource of Provenance-Centric Scientific. In Proceedings of IEEE eScience Linked Data (2018).
10. Heiko Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web 8(3): 489-508 (2017).
11. Cai, H., Zheng, V. W., Chang, & K. Chen-Chuan: A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. IEEE Transactions on Knowledge and Data Engineering (2017).
12. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven evaluation