Domain Specific Knowledge Graphs as a Service to the Public

Powering Social-Impact Funding in the US

Ying Li*
Giving Tech Labs
Seattle, United States
ying@giving.tech

Vitalii Zakhozhyi†
Giving Tech Labs
Seattle, United States
vitalii.z@giving.tech

Daniel Zhu†
Giving Tech Labs
Seattle, United States
daniel.z@giving.tech

Luis J. Salazar*
Giving Tech Labs
Seattle, United States
luis@giving.tech

ABSTRACT

Web and mobile technologies enable ubiquitous access to information. Yet, it is getting harder, even for subject matter experts, to quickly identify quality, trustworthy, and reliable content available online through search engines powered by advanced knowledge graphs. This paper explores the practical applications of Domain Specific Knowledge Graphs that allow for the extraction of information from trusted published and unpublished sources, to map the extracted information to an ontology defined in collaboration with sector experts, and to enable the public to go from single queries into ongoing conversations meeting their knowledge needs reliably. We focused on Social-Impact Funding, an area of need for over one million nonprofit organizations, foundations, government entities, social entrepreneurs, impact investors, and academic institutions in the US.

CCS CONCEPTS

• **Information systems** → **Graph-based database models**; **Data mining**; • **Computing methodologies** → **Information extraction**; **Ontology engineering**.

KEYWORDS

domain specific knowledge graph, social-impact funding, domain ontology

ACM Reference Format:
Ying Li, Vitalii Zakhozhyi, Daniel Zhu, and Luis J. Salazar. 2020. Domain Specific Knowledge Graphs as a Service to the Public: Powering Social-Impact Funding in the US. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394486.3403330

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403330

1 INTRODUCTION

Over 70 percent of the US population has access to the Internet using a mobile device [11], yet, it is getting harder to access accurate, timely, or trustworthy information, even for subject experts. Worldwide, citizens in 70 countries have been subject to political disinformation campaigns and social media manipulation [2]. In contrast to nonprofit organizations, academia, and government entities, the creators of misleading contents are often savvy users of social media and search engine optimization techniques deploying outsized budgets to increase content discovery. This asymmetry results in online content presenting a distorted or incomplete view of critical public issues, facilitating uninformed decisions, and potentially severe consequences.

In 2018, the nonprofit sector in the US received $427.71B in charitable donations [37], yet it is hard to answer questions like "who is funding research in Intellectual Disabilities?" or "who can I partner with to fund innovation that helps child survivors of abuse and neglect in the US?". General search engines often render incomplete information. Nonprofit sector experts agree that this lack of visibility results in inefficiencies in the deployment of charitable funding and government grants. It also affects the private sector, and tech entrepreneurs' interests in creating Technology for the Public Interest.

At Giving Tech Labs, we aim to create domain specific knowledge graphs as a service to the public. The goal is to extract the information from trusted published and unpublished sources, organize it under an ontology defined in collaboration with sector experts, and enable the public to go from single queries into ongoing conversations meeting their needs for accurate, reliable, trustworthy, and up-to-date information in topics that matter the most.

This paper focuses on the domain of **Social-Impact Funding**. By analyzing rich datasets from the Internal Revenue Service (IRS), the grants from the US Federal Government, and the funding activity of thousands of private foundations in the USA and the information available online of over one-million public charities, we can derive knowledge required to inform the future deployment of funds for social-impact. Our contributions are threefold:

(1) a domain ontology for Social-Impact Funding

† Work performed as *AI for Public Interest (AI4PI) Research Fellow* at Giving Tech Labs.
⋆ Funding was partially provided by Microsoft Technology for Social Impact division.

(2) a knowledge graph built for Social-Impact Funding
(3) insights derived from the knowledge graph to demonstrate its efficacy

The rest of this paper is organized as follows. Section 2 provides a brief overview of background knowledge for ontology and knowledge graph development. Section 3 introduces the ontology we developed for the Social-Impact Funding domain, and Section 4 provides details about the data and processes for building the domain specific knowledge graph. In Section 5, evaluation consideration is described and efficacy is demonstrated through a real-life use case. Finally, Section 6 concludes and outlines directions for future work.

2 BACKGROUND AND RELATED WORK

We are aware of that there are many different and even conflicting perspectives regarding a precise definition of a knowledge graph as pointed out by Ehrlinger et. al. [4]. For the purpose of this paper, we adopt the definition from Paulheim [33] and regard a knowledge graph to:

(1) mainly describe real world entities and their interrelations, organized in a graph;
(2) define possible classes and relations of entities in a schema;
(3) allow for potentially interrelating arbitrary entities with each other;
(4) cover various topical domains.

This characterisation of a knowledge graph is inclusive enough for us as practitioners to understand well-known industry scale knowledge graphs [30], such as the Google's Knowledge Graph [35] and Knowledge Vault [29], Microsoft's Academic Graph [38], etc. This "definition" of knowledge graph is also suited for us to cover the technical components in this paper while keeping some consistency with general references on the subject of knowledge graphs [8, 32] as well as with more recent research work on building knowledge graphs [22].

Although the distinction between general and domain specific knowledge graphs can be somewhat loose (Oxford dictionary defines a domain as "a specified sphere of activity or knowledge"), we consider domain specific knowledge graphs as those that model deeper knowledge, often only available from experts in the domains, and cover application issues in the specific domains. General knowledge graphs usually model entities and relationships in the world without domain restrictions and can be applied for solving various issues that does not require deep knowledge in a particular domain.

2.1 Domain Specific Knowledge Graphs

Recently, there has been research and development work on constructing domain specific knowledge graphs and using these knowledge graphs for solving problems in the corresponding domains, such as:

- Cybersecurity [19]
- Understanding the Impact of Opioid Crisis in the U.S. [20]
- Combating Human Trafficking [36]
- Factory Monitoring and Process Automation [12]
- Capturing Geological Data [40, 41]
- Geopolitical Events [21]

Taking the Siemens example, Hubauer et. al. [12] created a domain knowledge graph from multiple data silos within the enterprise. Based on the domain model codified into the knowledge graph, inference and machine processing are enabled which in turn enabled various applications such as factory monitoring, process automation, and others. Benefits brought about by domain specific knowledge graphs from the perspective of breaking and linking data silos can be captured at 4 levels [12]: 1) single data silos are enriched with domain specific models; 2) information flow between silos are enabled when data models for the silos are integrated; 3) an integrated data space is formed from the linked data silos when a tighter and more consistent integration between domains and use cases is established; and finally 4) an active knowledge factory is obtained by transforming the passive data space through integrating its vast amount of knowledge with graph-specific machine learning capabilities. The authors stated that for many of their use cases, knowledge graphs can already generate value on level 1 by facilitating information access for domain users.

There are also vendor/practitioner specific applications that utilize different aspects of knowledge graph related technologies, focused on solving a problem defined ahead of time for a specific domain. For such applications, often the data is already curated from enterprise data repositories. Sadowksi et. al. [34] uses graph database to discover connections to detect certain types of fraud. Winter [39] investigates drug repositioning through linking and mapping data schema. Such applications may not necessarily result in domain specific knowledge graphs that can be used to solve other and new problems in the same domain.

2.2 Domain Ontology Development

An ontology, a formal specification of concepts and relationships between them in a particular domain, captures the structure of data in the knowledge graph [24, 25, 31]. While there is no universal method to develop an ontology, the following design principles and criteria are widely used for the ontology construction:

- ***Resemblance to real world***: determine and scope its structure as close as possible to the real world [31];
- ***Fit for purpose and application***: optimize the structure for the intended purpose of the knowledge graph [17, 31];
- ***Iterative design***: start with a simple model and iterate with feedback from domain experts [25, 31];
- ***Scalability and extensibility***: support graph merging or sub-graph slicing for future growth in depth and extent [17].

Ontology development starts with defining the domain and scoping the ontology. Linková et. al. [25] refers to this stage as requirement analysis and suggests to consider ontology reusability and specificity as well as its coverage, richness , and cognitive adequacy. Noy and McGuinness [31] proposed to sketch a list of terms and concepts and a list of competency questions which the ontology should provide answers to determine the scope of the domain.

The next step of the ontology construction is the specification of the concepts and the relations between them. Within the defined scope of domain, terms can be grouped into entity types and attributes and relationships should be identified for each entity type in accordance with the potential applications of the ontology and the knowledge graph [31]. The initial ontology design should be

tested, evaluated, and refined through multiple iterations within a practical application.

2.3 Processes for Building Knowledge Graphs

A generic description for the process of a building knowledge graph consists of three main groups of tasks: 1) data and information extraction; 2) knowledge extraction; 3) knowledge serving.

The goal of ***data extraction*** is to extract information from different data sources, structured or unstructured, private or public, large or small. Typical tasks for data extraction include web crawling, entity and attribute extraction, and matching. The goal of ***knowledge extraction*** is to extract relationships between entities and infer attributes for entities and relations. Knowledge extraction tasks include relation extraction, concept and topic discovery, entity disambiguation, etc. ***Knowledge serving*** provides a mechanism for the user of the knowledge graph to construct their queries, execute the queries, and present the results for the user to derive insights.

Algorithms for those tasks exist and some are well established in the research areas of information extraction and retrieval and natural language processing [26]. Correspondingly, some of the algorithms are implemented and available as open source software or in commercial offerings [3, 5, 14]. However, even with the abundant availability of software and tools, human intervention is still important to ensure high coverage by the knowledge graph, and to improve the precision of machine learners.

The next section will introduce our ontology for Social-Impact Funding, then in subsequent sections we will present the process and system components we built for constructing the knowledge graph.

3 MODELING SOCIAL-IMPACT FUNDING

To develop the ontology for the domain of Social-Impact Funding, we applied the principles described in Section 2.2 and focused on two main steps that we will describe in detail: 1) defining and scoping the domain; 2) specification of entities and relationships in the domain.

3.1 Scope of Social-Impact Funding

The tasks of defining and scoping the Social-Impact Funding domain include: 1) creating a collection of concepts and identifying the relationships between them; 2) developing a list of competency questions; 3) modeling the impact according to the United Nations Sustainable Development Goals (SDG) [28].

The collection of concepts consists of the terms used within the Social-Impact Funding domain and specifies the possible relations between them. We built this collection by aggregating typical domain anecdotes from experts within the Social-Impact Funding domain and from resources such as news, academic, and popular science articles. We, then, extract entities (nouns) and relations (verbs) from these anecdotes.

We designed a set of competency questions based on the analysis of the Social-Impact Funding needs validated by representatives of the nonprofit, social entrepreneurship, academia, and impact investment sectors. Some of the resulting competency questions are:

- Who is funding what? Who funds subject **X** in a location **Y**?
- Who funds an organization **O** for more than two years?
- How does money flow for a social challenge or a subject **X**?
- Whom can an individual or an organization partner with to fund a social challenge or a solution to a pressing social problem?
- How much funding is allocated to each one of the SDGs?

Modeling impact to the SDGs amplifies the power of our ontology given the emerging consensus on aligning Social-Impact Funding with SDGs. The existence of impact can be determined by analyzing whether an activity has narrowed the gap between the existing state of art in some sphere (specific social issue) and the ideal future (SDG as a proxy). Adopted in the 2015 United Nation's master plan, SDGs laid out the thematic roadmap for the world development until 2030. This roadmap defines 17 comprehensive goals, 169 targets, and 232 indicators [28], which can be tracked nationally to evaluate the progress and contribution of each country to the global transformation towards a better and more sustainable world. We incorporated the SDGs' architecture of

$$Goal \Rightarrow Target \Rightarrow Indicator$$

in our ontology as a reference framework for social-impact measurement and standardization of positive contributions to solving social issues and investments into sustainable development [18].

To model the concept of impact in the ontology, we used the Outcome-Indicator idea from the *Theory of Change Model (Impact Value Chain)* [9, 10, 13]

$$Outcome + Indicator \Rightarrow Impact$$

meaning that accumulation of outcomes (short- and middle-term effects of organization's activities measured with specific indicators) over time result in the organization's impact towards solving a social issue.

3.2 Entities and Relationships Specification

After scoping the domain, we identified 3 larger facets which we want to focus on when constructing the ontology:

Funding sources - ***organization*** entity type represents both grantors and grantees with ***funds*** relationship; We also included ***geography*** entity type to capture the location of the organization and the area of its activities.

Social issues - ***subject*** entity type represents the organization's activity focus areas which can be extrapolated to a social issue it is working on. This relation can be static, such as the code the organizations registered with the IRS. It can also be progressively changing over time when the organizations shift their focus areas. In addition, we capture the beneficiaries of the organizations and their activities under ***beneficiary*** and ***beneficiary group*** entity types.

Impact/Outcome - ***outcome*** entity type contains information about results of an organization's activity. Through the ***indicator*** entity, we link organizations and their outcomes to specific ***SDG*** entity type as SDG framework is the standard blueprint for describing the measurement of progress towards solving social problems.

Our ontology for the Social-Impact Funding domain is presented in Figure 1. The entities and relations together allow us to address the competency questions raised in 3.1 by describing *who* provided

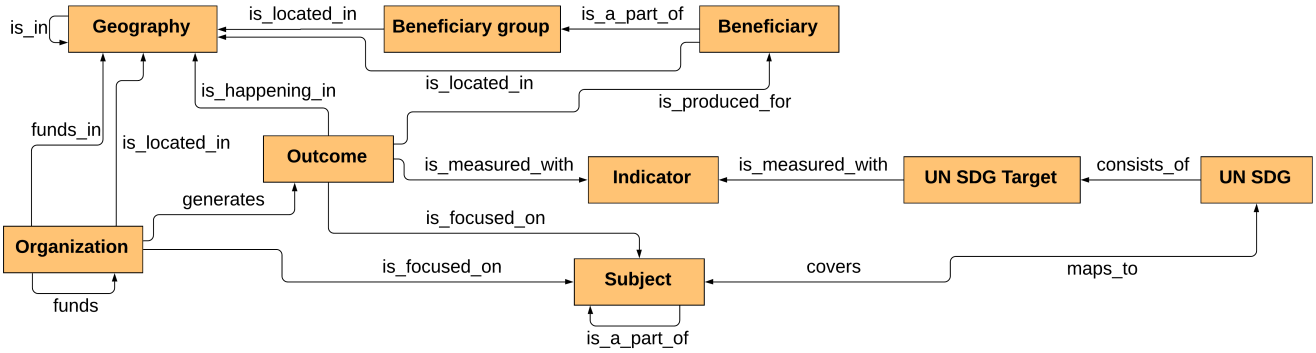


Figure 1: Ontology for the domain of Social-Impact Funding.

funding to *whom* and for *what* subject or for *which* beneficiary in *which* location and *how* those activities are mapped to SDGs.

4 CONSTRUCTION OF KNOWLEDGE GRAPH FOR SOCIAL-IMPACT FUNDING

This section describes the processes and the data utilized for constructing the knowledge graph according to the ontology described in previous section and Figure 1. Our overall architecture is presented in Figure 2.

The data sources that are input to the system need to provide trustworthy and comprehensive coverage on Social-Impact Funding domain. Harvesting the entire web for data and knowledge about Social-Impact Funding is not feasible because the signal is extremely sparse on the general web and is continuously being drowned out by noise. Thus, we decided to start with trusted data sources that are curated from domain experts and reliable partners and institutions, detailed in Section 4.1.

Sections 4.2 - 4.5 will describe details of the transformation process from the raw data to entities and relations as specified by the domain ontology. The system is an ensemble of components commonly used for processing and extraction of knowledge from unstructured text data; however, there are data and technical nuances that, if not handled properly, could impact the quality of the final knowledge graph in material ways. We focus on our solution to some of these nuances in the following sections.

4.1 Trusted Data Sources

It is known from research in knowledge graphs that the completeness of a knowledge graph is an untenable goal. Hence, our strategy for domain data and knowledge is to 1) strive for 100% trustworthiness by partnering with domain experts, and 2) strive for statistical validity and usefulness in generating domain insights usable by domain experts and the general public.

To build the domain knowledge graph according to the ontology described in Figure 1, we use the following sources:

- official or verified websites for:
  - nonprofit organizations and community initiatives
  - public and private foundations
  - US government agencies
- portals and aggregators for funding and grants such as grants.gov, a list all federal grants
- organizations' tax returns (a statistically significant subset from 2013 to 2019 available on AWS [16])

To facilitate impact modeling to SDGs, a set of articles and white-papers for each of the 17 SDGs are manually curated from the United Nations and organizations that focus on SDGs. These documents detail specific information about the respective SDG. Each document was labelled with the respective SDG by domain experts. The data set serves as ground truth for building domain dictionaries and impact classifiers, and we refer to it as **SDGTrain** data set.

4.2 Entities and Relations from IRS Data

The Internal Revenue Service (IRS) is a rich and trustworthy data source for uncovering the intricacies of Social-Impact Funding. The transformation from the raw IRS 990 data to entities and relations in the ontology (Figure 1) consists of two steps:

(1) **extraction** - entities, relations, and their corresponding attributes are exact-matched from data fields, producing accurate but not complete results
(2) **inference** - interpolate missing relations and entity attributes by utilizing contextual information and knowledge graph assisted reasoning

To build the initial structure of the knowledge graph, we start with the extraction of ***organization*** and ***geography*** entities. We first sourced a complete list of ***organization*** entities from the IRS Business Master File and indexed by the Employee Identification Number (EIN). However, due to structural nuances such as mergers, acquisitions, organizational chapters, among others, an organization can have multiple EINs. We resolve these issues by consolidating the equivalent entities into one ***organization*** entity. Then, we create ***geography*** entities to cover the state and zip-code level designations within the US and utilize the relation ***is_in*** to determine the hierarchy.

From each 990PF e-filing, we extract the relation ***is_located_in*** between an ***organization*** and ***geography*** at the zip code level. A small percent of ***is_located_in*** relations are removed pre-ingestion due to observing human error in the 990PF filing.

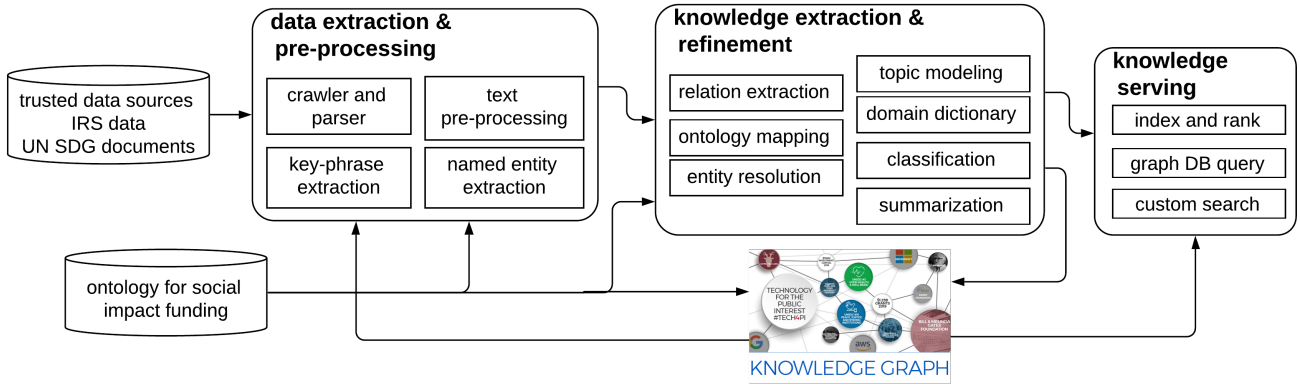


Figure 2: General description of the main tasks involved in building a knowledge graph.

The relation $A \xrightarrow{funds} B$ takes a combined approach. ***Organization A***, the grantor, can by matched by the unique EIN specified in the 990PF form entry. Then, ***organization B*** is retrieved by matching the given name and the address information of grantees. However, we found that the entity matching can fail on both ***A*** and ***B***. For ***A***, a historical issue arises, the ***organization*** entities represent a current collection of tax-exempt organization recognized by the IRS, so an organization that loses its tax-exempt status in past years would not be included. More complicated is ***B***. Consider the following example from the Bill & Melinda Gates Foundation's 2017 990PF e-filing where they have listed the following grant:

Gates Foundation $\xrightarrow{funds}$ **Heroes for the Homeless**

EIN: 91-1663695

EIN: ?
Street: PO BOX 418
City: Woodinville
State: WA
Zipcode: 98072

To map this data, we complete soft matching with existing ***organization*** entities in the knowledge graph on the cleaned up contextual information of the grantee, providing a reliable match if one exists. Again, we suffer from the historical challenge: changing addresses and names through the years. ***B*** should match to the following ***organization*** that has been updated for 2019:

Heroes for the Homeless
EIN: 20-8634738
Street: 10562 NE 122nd St.
City: Kirkland
State: WA
Zipcode: 98034

An additional concern is that the grantee as appears in the 990PF form is not guaranteed to be of entity type ***organization***; grants are often for individual recipients or the result of employee matching. To ensure accuracy, we remove these instances from the graph.

Once a match is found for ***organization B***, we enrich the Social-Impact Funding Knowledge Graph with the ***funds*** relation between ***organization A*** and ***organization B***. After ingesting the entire collection of ***funds*** relations from the IRS data, we further discover the relation ***funds_in*** by inferring from the following path:

$$A \xrightarrow{funds} B \xrightarrow{is\_located\_in} geography$$

therefore,

$$A \xrightarrow{funds\_in} geography$$

From the IRS 2013 - 2019 filing indices, we extracted a partial, yet extensive picture of the funding activity throughout these years. The comprehensiveness of the data was measured by the magnitude of entities and relations in Table 1. Note also this table also contains the number of dictionary words per SDG which we will present at next section.

4.3 Dictionary for Social-Impact Funding

A domain specific dictionary strives to understand the dominant concepts within the domain. We applied the approach from Kim et. al. [23] to build our SDG domain dictionary. The results of scraping all sources from **SDGTrain** dataset at the paragraph level were conflated within each SDG to produce a representative domain level document for each SDG. Using this corpus, we calculate the domain TF-IDF specifying each SDG and output the top 100 words based on TF-IDF score. We repeat this procedure for bigrams and trigrams and aggregated the resulting domain dictionaries to form our eventual SDG domain dictionary. The size of the resulting dictionary is shown in Table 1 (column 4). We use domain experts to validate the quality of the dictionary, and prune false positives when necessary.

An excerpt from the domain dictionary for SDG #16: Peace, Justice, and Strong Institutions is as follows:

> *freedom_expression, bribery, activism, criminal, fundamental_freedoms, participatory_representative, strengthening_institutions, peacebuilding, corruption, access_justice, peace, judicial, freedom, ...*

Take the terms *freedom* and *corruption* from the above excerpt, these words can be found in varied contexts in the general sense. But, within the SDG scope, we surmise that word distance between these terms within the scope of SDG domain would differ from

Table 1: Scope of matched entities and relations contained in Social-Impact Funding Knowledge Graph from IRS sources.

United Nations Sustainable Development Goals	Organization Entities	Funding Relations	Dictionary Terms
SDG #1: No Poverty	40,336	29,881	61
SDG #2: Zero Hunger	31,186	30,480	85
SDG #3: Good Health and Well-being	218,074	170,835	73
SDG #4: Quality Education	192,285	270,025	62
SDG #5: Gender Equality	34,675	20,534	68
SDG #6: Clean Water and Sanitation	4,675	9,562	78
SDG #7: Affordable and Clean Energy	13,208	20,310	69
SDG #8: Decent Work and Economic Growth	49,326	18,958	92
SDG #9: Industry, Innovation, and Infrastructure	36,800	11,802	69
SDG #10: Reduced Inequality	122,264	94,972	83
SDG #11: Sustainable Cities and Communities	112,413	63,701	88
SDG #12: Responsible Consumption and Production	1,311	371	63
SDG #13: Climate Action	33,370	34,770	79
SDG #14: Life Below Water	10,350	19,168	84
SDG #15: Life on Land	47,214	67,198	86
SDG #16: Peace, Justice, and Strong Institutions	77,729	50,905	100
SDG #17: Partnerships to achieve the Goals	21,631	22,947	*18
Total	**1,046,847**	**936,419**	**1258**

* SDG #17 has few domain specific words because it overlaps with the implementation of the other goals and employs the same terminology.

the general usage due to the increased co-occurrence when discussing SDG #16. We verify this hypothesis by comparing a generic Word2Vec [27] to a domain specific Word2Vec, which we train using the **SDGTrain** dataset to capture the contextual differences of words when discussed under the SDG framework.

SDG domain Word2Vec is used to reinforce the relationship of dictionary terms within the SDG space. Performing PCA on word embeddings of the dictionary terms in each SDG, we can confirm the validity of dictionary terms in the SDG domain dictionary by observing the clustering of terms within each SDG. Figure 3 shows one visualization of dictionary terms for SDG #16. The combination of a domain specific dictionary and Word2Vec enables us to perform the downstream task of SDG classification that we will discuss next.

4.4 Impact Classification

To analyze the impact of funding, we align the involvement of ***organization*** to ***SDG*** through the following path of the ontology:

$$\textit{Organization} \xrightarrow{\textit{is\_focused\_on}} \textit{Subject} \xrightarrow{\textit{maps\_to}} \textit{SDG}$$

For an ***organization*** that has an attached ***subject***, such as National Taxonomy of Exempt Entities (NTEE) classification [6] from IRS, we can create a mapping from ***subject*** to ***SDG***. For the cases of NTEE as subject, we adopted a one-to-many mapping from NTEE codes to SDGs designed by domain experts.

When the ***is_focused_on*** relation cannot be readily extracted, text classification methods are used on the context of ***organization*** to discover such relations. We built multi-class text classifiers

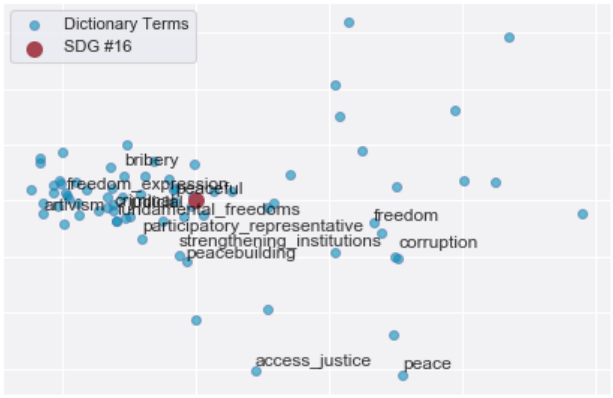


Figure 3: PCA visualization of the distance of dictionary terms from SDG #16 based on a domain Word2Vec.

with labeled training data from the **SDGTrain** paragraph-level documents. We perform the text cleaning steps of punctuation and stopword removal and trigram transformation.

To evaluate the generalization error of the trained classifier, we tested the ability of the classifier to classify NTEE definitions and computed the accuracy against our NTEE to SDG mapping. We used Top-3 evaluation metric out of two considerations: 1) there exists a one-to-many relation from NTEE code to SDG; and 2) social activities can be involved in multiple SDGs. Top-3 evaluation can also account for individual organizational differences ignored by the NTEE to SDG mapping.

Our first impact classifier, we trained a multi-layer perceptron on the TF-IDF feature space [15]. Using the Top-3 criteria, this model achieved a **86%** accuracy on the validation set, a partition of the training set, and **78.8%** accuracy on the NTEE test set. While multi-layer perceptrons have performed well in practice for text classification, the effects of small size of our the **SDGTrain** dataset and imbalance between classes were observable in the results: minority classes, the SDGs with less rich and a lower quantity of documents, were misclassified into the majority classes.

Faced with insufficient training data, we proposed a text classification algorithm based on keyword similarity matching to a SDG domain. We transform each word document into the set of its word embeddings. For a set of n dimensional word embeddings, the score vector, detailing the similarity score to each class, for a document is produced by the following:

$$S(\mathbf{d}) = \frac{\mathbf{c}^{\mathrm{T}}\mathbf{C}\mathbf{d}}{\|\mathbf{d}\|_2}$$

where $\mathbf{C} \in \mathbb{R}^{dxn}$ is matrix where the ith row vector corresponds to the centroid of class i. Then, $\mathbf{c} \in \mathbb{R}^d$ is the vector containing the L2 norms of each class centroid in order. $\mathbf{d} \in \mathbb{R}^n$ is the average of the word embeddings of the keywords within a document.

Applied to the impact classification problem, the algorithm operates on the basis that document's representation can be derived from the average of the word embeddings of its keywords and similarly a SDG centroid can be represented by the average of the word

embeddings of the dictionary terms. We then measure the relatedness of a document to each SDG class by calculating the cosine similarity between the document vector and the SDG centroid. The accuracy of the model is heavily dependent on the robustness of it's components:

(1) *Domain Specific Dictionary* - determines the centroid of each class
(2) *Domain Word2Vec* - adjusts the measure of similarity for domain specificity

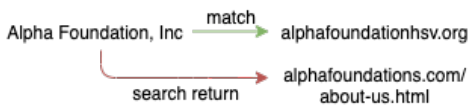
The resulting model utilizes the SDG domain dictionary and SDG Word2Vec described in previous section. On our NTEE test data set, we achieved an accuracy of **36.7%**. An analysis of the misclassifications showed the inability to classify select SDGs, we aim to improve the performance of the classifier by investigating the respective domain dictionaries.

4.5 Web Data Extraction

To be able to apply a text classification solution to the inference task of an organization's involvement in the 17 SDGs requires a representative document describing the activities of an organization. Often a fitting document to detail an organization's activities is their "mission", "vision", or "about" statement. This information is not available on the IRS tax related forms. Therefore, we expanded our search to the unstructured web to enrich the contextual information for each organization.

For the purpose of expanding the context of an organization, an apparent resource is *Wikipedia*. We linked organizations to *Wikipedia* pages using Microsoft Text Analytics entity extraction [3]. The results reveal a core issue: the lack of a Wikipedia presence for private foundations. Only **4%** matched *Wikipedia* pages were reasonable - that is, the returned *Wikipedia* page is an exact match or a broader representation of the foundation.

To address this issue, we broadened our coverage by using Microsoft Bing Search service to perform a search for a organization's web page, particularly the "about" page. For each organization, we narrowed the search results using a criteria that searched for URLs that contained the greatest number of words in the organization's name and contained "about", "vision", or "mission" in the title. This method yielded a more bountiful return on organization URLs. After applying a filter for URLs containing "about", **46%** of the matched URLs were appropriately matched to the foundation. Even then, the underlying assumptions from the search criteria are easily invalidated. Consider this case of matching:



The search return belongs to a foundation repair company that does not fall under the scope of our domain, not the intended nonprofit Alpha Foundation. When limited to string matching the URL and metadata, the search return contends with the matching URL. To avoid polluting the knowledge graph with mismatched web data, we address another text classification problem to determine whether text content is relevant to our specific domain or not.

We created a SDG/non-SDG dataset from the NTEE mapping by considering all NTEEs mapped to a SDG as domain content and those unable to be mapped as out of domain content. Through a seed plus expansion method where each NTEE definition serves as a seed query, we expand the dataset through supplying the definition as a query to Google Search engine and scraping the top 3 web pages to provide additional training data. After text preprocessing, we train a Naive Bayes model, which performs at a **2%** false positive rate when tested on a subset of web scraped organization descriptions. We apply this binary classifier as an initial filter to optimize the specificity and trustworthiness of the ingested data.

Using a subset of reasonably matched organizations, we enhanced the descriptions of the organizations by scraping the matched web page. We classify on this subset of organizations and the results were reviewed manually. **77.2%** of the organizations were categorized into a reasonable SDG affiliation. The Social-Impact Funding Knowledge Graph was then enriched with the ***is_focused_on*** from classification results.

5 EVALUATION AND APPLICATIONS

A knowledge graph can be evaluated [1, 33] in multiple aspects such as accuracy, trustworthiness, consistency, relevancy, completeness, timeliness, ease of understanding, interoperability, accessibility, licensing, and interlinking, etc. However, researchers agree that it is challenging to simultaneously improve the quality of knowledge graphs in many different aspects; it is an open question of what aspects are most critical and necessary. We focus on trustworthiness through selecting reliable data sources trusted by domain practitioners. We focus on efficacy through testing the knowledge graph for solving real-life problems and assessing testimonies from domain experts. Also, for accuracy, we evaluate specific data points against known data sources.

5.1 Evaluation of Data Accuracy

The evaluation of a knowledge graph is a continuous maintenance effort, and our Social-Impact Funding Graph is no exception. We have identified specific metrics to serve as evaluation benchmarks and audited distinct data points and validating them using publicly available data and internal data from organizations.

We used the total size of the nonprofit sector in the US to confirm a validity of our data extraction from IRS forms. The Social-Impact Funding Knowledge Graph gives us **$2.12 trillion** reported by US nonprofits over the last 12 months, which matches the generally accepted size of the nonprofit sector of **$2 trillion** [7]. Another metric we used was a total amount of charitable giving in the total income of the US nonprofit sector, which is estimated of **$427.71 billion** in 2018 by Giving USA [37]. This matched the findings in our knowledge graph of **$450.7 billion**.

5.2 Use Cases to Demonstrate Efficacy

We consulted with subject matter experts to understand how the lack of access to insights had a detrimental impact on the funding of social-impact causes. A Program Manager Officer at the Raikes Foundation and a leader at a Community Foundation with $1B under management, had similar needs to understand who else they can partner with to support a specific cause, a group of grantees

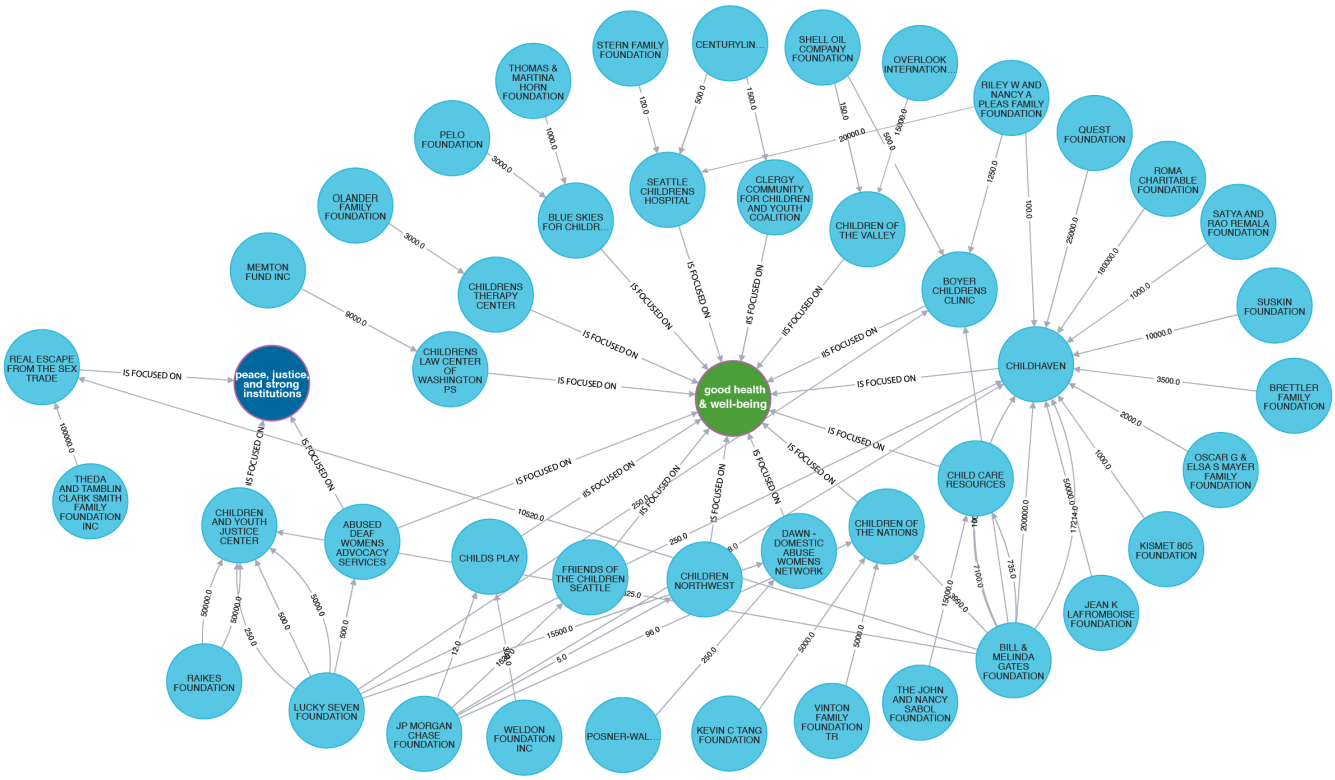


Figure 4: Visualization of a sub-graph derived from our Social-Impact Funding Knowledge Graph. The sub-graph represents over $850,000 of social-impact funding focused on children's justice and health in WA.

or an individual organization. For example, in the case of a funder who is looking for funding partners who supports programming activities related to criminal justice and healing for survivors of child abuse and neglect in WA state, they would use the following query in a conventional search engine:

```
Funding AND Justice AND Health AND WA AND
(Abuse OR Child OR Sex OR Neglect)
```

This search yields about 7,380,000 results, the top of which provides useful links related to abuse and neglect on children, but no information related to funding flow. The question remain unanswered despite of the use of the complex search query.

We use Neo4j as our graph database that enables us to store and query the knowledge graph. We structure our query in Neo4j for the above question as follows:

```
JUSTICE and HEALTH in WA
MATCH q=(Donor:Organization)-[fund:FUNDS]->
(Org:Organization)-[:INVOLVED_WITH]-(un:SDG)
WHERE ((un.name CONTAINS "justice") OR
(un.name) CONTAINS "health")
AND NOT fund.purpose CONTAINS "MATCHING"
AND (fund.tax_period) = 2017
AND (Org.state) = "WA"
AND ((Org.name) CONTAINS "ABUSE" OR
(Org.name) CONTAINS "CHILD" OR
(Org.name) CONTAINS "SEX"
OR (Org.name) CONTAINS "NEGLECT")
RETURN q
```

The resulting visualization displayed in Figure 4 helped informed the decision in the following cases:

(1) A program officer of a private foundation connected with a peer foundation that also invests in justice for children, to plan an impact-funding initiative with a new grantee.
(2) A Director of VidaNyx, a digital video evidence management solution that serves the needs of Child Advocacy Centers, identified relationships between a client of VidaNyx and several grantors that help other Child Advocacy Centers, which are still using manual processes. Now she can pursue social-impact funds from those grantors to support technology for those centers.

This demonstrates the advantage of a domain specific knowledge graph.

6 CONCLUSIONS AND FUTURE WORK

Domain-specific knowledge graphs can provide access to reliable and actionable data insights. As society continues to consume more content online, the creation of domain specific knowledge graphs, under ethical considerations, can create a structural shift in how nonprofits, foundations, government agencies, social entrepreneurs, impact investors, academic institutions, and the general public inform their actions towards creating social impact. This is showcased by the promising results validated by subject matter experts utilizing the Social-Impact Funding Knowledge Graph.

We can apply the same methodology developed for the construction of the Social-Impact Funding domain graph to other domains. We organized the existing knowledge on the topic and developed the reproducible methodology for both ontology and knowledge graph construction. In this paper, we laid out the challenges we encountered during this work, potential ways to overcome them, and the lessons we learned.

As a next step, we plan to use the methods and system we described in this paper to enhance the Social-Impact Funding Knowledge Graph. We plan to add complementary datasets from the Security and Exchange Commission (SEC), the Venture Capital and Private Equity industries, and the impact investing sector, to derive greater knowledge about the total flow of funding towards social-impact causes. Another step is to use the methodology to develop a knowledge graph for the intellectual disabilities domain in partnership with Special Olympics International.

REFERENCES

[1] Piero A. Bonatti, Michael Cochez, Stefan Decker, Axel, Polleres, and Valentina Presutti. 2018. Knowledge Graphs : New Directions for Knowledge Representation on the Semantic Web. In *Dagstuhl Seminar 18371*. https://www.dagstuhl.de/18371

[2] Samantha Bradshaw and Philip N. Howard. 2019. The Global Disinformation Disorder: 2019 Global Inventory of Organised Social Media Manipulation. https://comprop.oii.ox.ac.uk/research/cybertroops2019/

[3] Microsoft Corporation. 2019. Text Analytics API documentation. https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/

[4] Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a Definition of Knowledge Graphs. In *SEMANTiCS*.

[5] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. 2014. Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems* 70 (Nov 2014), 301–323. https://doi.org/10.1016/j.knosys.2014.07.007

[6] National Center for Charitable Statistics of the Urban Institute. 2019. National Taxonomy of Exempt Entities (NTEE) Codes. https://nccs.urban.org/project/national-taxonomy-exempt-entities-ntee-codes

[7] National Center for Charitable Statistics of the Urban Institute. 2019. The Nonprofit Sector in Brief. https://nccs.urban.org/project/nonprofit-sector-brief

[8] Yuqing Gao, Jisheng Liang, Benjamin Han, Mohamed Yakout, and Ahmed Mohamed. 2018. Building a Large-Scale, Accurate and Fresh Knowledge Graph. https://kdd2018tutorialt39.azurewebsites.net/KDD%20Tutorial%20T39.pdf

[9] M.K. Gugerty and D. Karlan. 2018. *The Goldilocks Challenge: Right-fit Evidence for the Social Sector*. Oxford University Press. https://books.google.com/books?id=6qZTDwAAQBAJ

[10] Dr Lisa Hehenberger, Anna-Marie Harling, and Peter Scholten. 2015. *A Practical Guide to Measuring and Managing Impact*. Technical Report. European Venture Philanthropy Association. https://evpa.eu.com/knowledge-centre/publications/measuring-and-managing-impact-a-practical-guide

[11] Arne Holst. 2019. Smartphone Penetration in the US 2010-2021. https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/

[12] Thomas Hubauer, Steffen Lamparter, Peter Haase, and Daniel M. Herzig. 2018. Use Cases of the Industrial Knowledge Graph at Siemens. In *International Semantic Web Conference*.

[13] Ilma Ibrisevic. 2018. Measuring Nonprofit Social Impact: A Crash Course. https://donorbox.org/nonprofit-blog/measuring-nonprofit-social-impact/

[14] Google Inc. 2019. AutoML Natural Language. https://cloud.google.com/natural-language/#overview

[15] Google Inc. 2019. Google Machine Learning: Text Classification. https://developers.google.com/machine-learning/guides/text-classification/step-2-5

[16] IRS. 2016. IRS 990 Filings on AWS. https://registry.opendata.aws/irs990/

[17] Wątróbski Jarosław. 2018. An Attempt to Knowledge Conceptualization of Methods and Tools Supporting Ontology Evaluation Process. *Procedia Computer Science* 126 (2018), 2238 – 2247. https://doi.org/10.1016/j.procs.2018.07.225 Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia.

[18] Mark Jensen. 2016. Sustainable Development Goals Interface Ontology. In *ICBO/BioCreative*. https://github.com/SDG-InterfaceOntology/sdgio/tree/master/docs/term%20lists

[19] Yan Jia, Yulu Qi, Huaijun Shang, Rong Jiang, and Aiping Li. 2018. A Practical Approach to Constructing a Knowledge Graph for Cybersecurity. *Engineering* 4, 1 (2018), 53 – 60. https://doi.org/10.1016/j.eng.2018.01.004 Cybersecurity.

[20] Maulik R. Kamdar, Tymor Hamamsy, Shea Shelton, Ayin Vala, Tome Eftimov, James Zou, and Suzanne Tamang. 2019. A Knowledge Graph-based Approach for Exploring the U.S. Opioid Epidemic. arXiv:cs.CY/1905.11513 https://arxiv.org/abs/1905.11513

[21] Mayank Kejriwal. 2019. *Domain-Specific Knowledge Graph Construction*. Springer. https://doi.org/10.1007/978-3-030-12375-8

[22] Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. In *AAAI Workshops*.

[23] Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. 2009. Extracting Domain-Specific Words - A Statistical Approach. In *Proceedings of the Australasian Language Technology Association Workshop 2009*. Sydney, Australia, 94–98. https://www.aclweb.org/anthology/U09-1013

[24] Agnieszka Konys. 2018. Knowledge systematization for ontology learning methods. *Procedia Computer Science* 126 (2018), 2194 – 2207. https://doi.org/10.1016/j.procs.2018.07.229 Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia.

[25] Zdeňka Linková, Radim Nedbal, and Martin Rimnac. 2005. Building Ontologies for GIS. (01 2005).

[26] Christoper Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*.

[27] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (2013). http://arxiv.org/abs/1301.3781

[28] United Nations. 2015. The United Nation Sustainable Development Goals. https://sustainabledevelopment.un.org/sdgs

[29] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proc. IEEE* 104, 1 (Jan 2016), 11–33. https://doi.org/10.1109/JPROC.2015.2483592

[30] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-Scale Knowledge Graphs: Lessons and Challenges. *Commun. ACM* 62, 8 (July 2019), 36–43. https://doi.org/10.1145/3331166

[31] N. Noy and Deborah Mcguinness. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. *Knowledge Systems Laboratory* 32 (01 2001).

[32] Jeff Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu. 2017. *Exploiting Linked Data and Knowledge Graphs in Large Organisations*. https://doi.org/10.1007/978-3-319-45654-6

[33] Heiko Paulheim. 2016. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* 8 (2016), 489–508.

[34] Gorka Sadowksi and Philip Rathle. 2017. Fraud Detection: Discovering Connections with Graph Databases. https://go.neo4j.com/rs/710-RRC-335/images/Neo4j_WP-Fraud-Detection-with-Graph-Databases.pdf?_ga=2.152229817.1435723348.1577409683-120002542.1565112145

[35] A. Singhal. 2012. Introducing the Knowledge Graph: Things, Not Strings. http://goo.gl/zivFV

[36] Pedro Szekely, Craig A. Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, David Stallard, Subesswara S. Karunamoorthy, Rajagopal Bojanapalli, Steven Minton, Brian Amanatullah, Todd Hughes, Mike Tamayo, David Flynt, Rachel Artiss, Shih-Fu Chang, Tao Chen, Gerald Hiebel, and Lidia Ferreira. 2015. Building and Using a Knowledge Graph to Combat Human Trafficking. In *The Semantic Web - ISWC 2015*. Springer International Publishing, 205–221.

[37] Giving USA. 2019. Giving USA 2019: The Annual Report on Philanthropy for the Year 2018. https://givingusa.org/giving-usa-2019-americans-gave-427-71-billion-to-charity-in-2018-amid-complex-year-for-charitable-giving/

[38] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019. A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data* 2 (2019), 45. https://doi.org/10.3389/fdata.2019.00045

[39] Andrew Winter. 2019. Drug Repositioning Investigation Workflow on a "Virtualized" Knowledge Graph. https://siren.io/drug-repositioning-investigation-on-virtualized-knowledge-graph/

[40] Mingxiong Zhao, Han Wang, Jin Guo, Di Liu, Cheng Xie, Qing Liu, and Zhibo Cheng. 2019. Construction of an Industrial Knowledge Graph for Unstructured Chinese Text Learning. *Applied Sciences* 9, 13 (2019). https://doi.org/10.3390/app9132720

[41] Yueqin Zhu, Wenwen Zhou, Yang Xu, Ji Liu, and Yongjie Tan. 2017. Intelligent Learning for Knowledge Graph towards Geological Data. *Scientific Programming* 2017 (02 2017), 1–13. https://doi.org/10.1155/2017/5072427