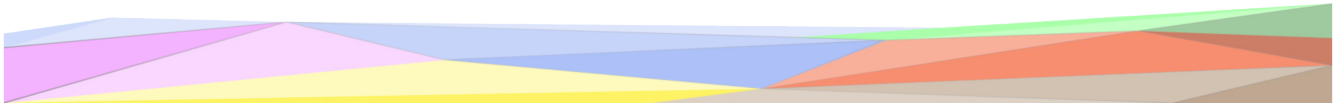
UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2022 - Metadata Project Report

Document Data:	Reference Persons:
November 14, 2022	Pelagalli Camilla, Alessandro Salvatore Raho



Index:

1 Introduction 1

2 Purpose and Domain of Interest (DoI) 1

3 Data Sources 2

4 Purpose Formalisation 3

5 Inception 6

Revision History:

Revision	Date	Author	Description of Changes
0.1	20.04.2020	Fausto Giunchiglia	Document created
0.2	13.11.2020	Pelagalli, Raho	Inception Phase added

1 Introduction

Reusability is one of the main principles in the Knowledge Graph Engineering (KGE) process defined by iTelos. The KGE project documentation plays an important role in order to enhance the reusabiltiy of the resources handled and produced during the process. A clear description of the resources and the process developed, provides a clear understanding of the KGE project, thus serving such an information to external readers in order to exploit that in new projects.

The current document aims to provide a detailed report of the KGE project developed following the iTelos methodology. The report is structured, to describe:

- Section 2: The project's purpose and the domain of interest and the resources involved (both schema and data resources) in the integration process.
- Section 3: The input resources considered by the KGE project.
- Section 4, 5, 6, 7: The integration process along the different iTelos phases, respectively.
- Section 8: How the result of the KGE process (the KG) can be exploited.
- Section 9: Conclusions and open issues summary.

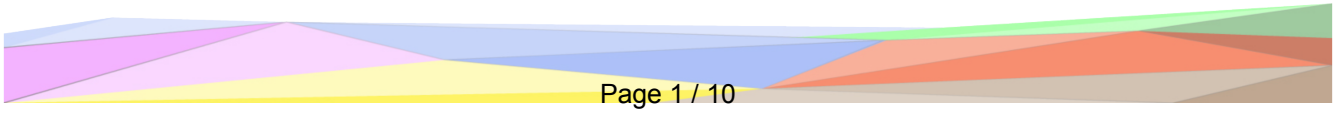
2 Purpose and Domain of Interest (DoI)

Our project context is an **Open Data environment**, where anyone is free to use, re-use or redistribute resources - data or contents - while preserving provenance and openness.

Data Scientia is the concrete representation of this environment: an under-costruction web portal that allows users to explore different kind of datasets. The portal contains two main web pages for now:

- **Homepage**: the research group's Vision and Mission are outlined and the main initiatives are globally described and referenced, such as UKC, SCroLL, Digital University and Smart University (http://datascientia.disi.unitn.it/).
- **Liveschema**: a catalog of knowledge resources, related to single projects. The user can retrieve different resources singularly, by searching inside Catalogs or by browsing datasets directly. There are 7 Catalogs (DERI Vocabularies, FINnish Thesaurus and Ontology service, GitHub Repository, KnowDive Vocabularies, Linked Open Vocabulary, Research Vocabularies Australia and User defined Datasets) and 958 Datasets. The user can filter datasets by Providers, Tags, Formats and Licenses. Available Services are: Dataset uploader, Knowledge embedder, FCA generator, Cue generator, Visualization generator and Query Catalog (http://liveschema.eu/).

Data Scientia's main goal is to gather all **projects** that result from Knowledge Graph Engineering (KGE) processes following the iTelos methodology ([Giu+21]). Each project can contain multiple



datasets, each belonging to a different informational layer. For each project, three main layers should be identified:

- **Language Layer**: any linguistic resource - property or dataset about languages and terms used - should be explicitly referenced to increase re-usability.
- **Knowledge Layer**: any Knowledge Base encoding information about KG schemas (etypes and properties). If detailed enough, this layer could be decomposed in its core elements: ontology and teleology, thus giving the user an unprecedented insight into the conceptual and relational data shaping behind the actual data values representation.
- **Data layer**: any dataset which consists of data in some format, instantiating the KG's structure (entities and attributes). If users are interested in expanding data resources, they can use the already existing schemas as meaningful anchors, without loosing data quality.

Data Scientia users should have access to both projects and their layers and decide what to re-use, modify or produce to meet the purpose they have in mind. Projects should be linked based on their similarities on a general and layered basis.

Metadata play a crucial role in this context because they describe datasets and represent the connections between them, making resources findable and addressable without any effort. For example, users looking for healthcare ontologies should be able to access them directly, without having to go through whole projects at the risk of stepping away from their purpose. Metadata reduces search time and space: they allow users to skip to what is essential to them, making the Data Scientia successful and meaningful.

Our project **purpose** can now be stated unambiguously:

*A KG supporting the **Data Scientia** web portal's users to find the most suitable **resource** for their needs, as well as all the resources **linked** by the one searched.*

To achieve the project objective, we have analyzed existing KGE Projects and Data Scientia resources and projects, which were our **data resources**. We have extracted the fundamental layered view of datasets from KGE Projects and we have outlined Data Scientia main entity types, along with their properties, by carefully examining the web portal and through a Data Scientia top level description and requirements given to us by the domain expert. The latter helped us understanding the preliminary boundaries of our **schema resources**, by stating the present structure of the portal with its shortcomings: the absence of linked resources and informational layers. To overcome them, we have designed a possible and plausible future structure for the Data Scientia web portal, by defining its Knowledge Graph.

3 Data Sources

As a starting point for our **Knowledge sources**, we have discussed Data Scientia inner structure with the domain expert. As shown in Figure 1, the resulting basic entity types of the skeleton

portal are the following:

- **Data Scientia**: the root portal, containing all the resources we need to manage, manipulate and integrate to meet the external user's needs and requirements.
- **Webpage**: the user-friendly interface(s) where resources are loaded and user can make queries and browse available services.
- **Person**: Data Scientia is a community and community members are assumed to interact directly with the web portal. A Person entity type is crucial to understand the boundaries between different roles of an individual with respect to the portal. The user should also have an easy way to find information about community members, partners, authors, researchers and so on. For example, these personas need to be described through the Organization they work for, or their publications. From now on we will name this group of people Data Scientia community.
- **Project**: the core entity type we have to consider, along with its Purpose. Projects can be uploaded to - or downloaded from - the portal and they also need to be decomposable and recomposable as KGE processes, according to their informational layers.

Starting from this draft, since we currently do not have access to any metadata datasets, we have collected data resources from Data Scientia already existing web pages and thought about which kind of new web resources would be useful to add or integrate. We have read the documentation about existing projects, uploaded to both the Data Scientia portal and the liveschema catalog, along with KGE projects which followed the iTelos methodology. We have explored their components carefully to extract the fine-grained metadata used to create the property columns of the attached e-types labels. We have also focused on which metadata are used in DataScientia to depict people.

4 Purpose Formalisation

We have defined **6 scenarios** and matching **personas**, paying attention to the different resources relative to a specific KGE process. Users should be able to look at the informational layers separately and as a whole project when browsing the web portal.

The **first** scenario describes the activity of navigating resources linked to a specific purpose. The corresponding personas are KGE course students. They need to browse already existing KGE projects to understand assignments and also check if previous projects tackled their purpose or have a similar description or theme. Students could re-use the ontology to manage and integrate data or have access to datasets they don't know where to find.

The **second** scenario is about general searching of datasets: personas want to check only data layers of different projects, their purpose is to download as much data as possible, regardless of the projects purpose or common themes. They want to perform operations on actual data values, such as filtering, querying, pattern matching, which not necessarily involve meaningful treatment of existing information.

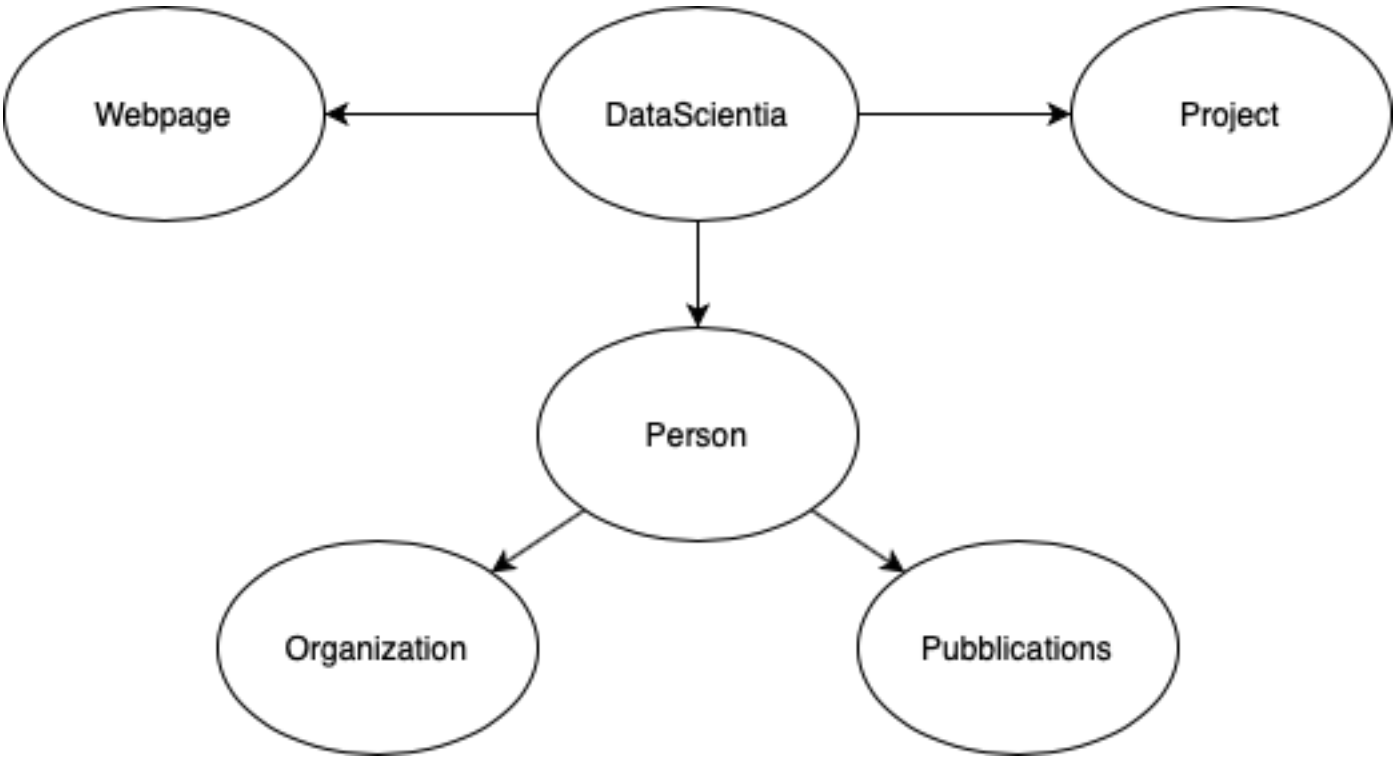


Figure 1: DataScientia Structure

The **third** scenario explains the possibility of making resources available under a specific license. Personas could be university researchers interested in sharing their projects and getting recognition. They may want to upload single projects layers or datasets linked to a well-defined context. Reachability is their main concern, the produced resources need to reference directly their personas and other existing projects addressing the same issue, or theme, should be linked to theirs to increase popularity and public coverage.

The **fourth** scenario depicts the action of looking for a specific informational layer, for example the knowledge layer, according to a personal purpose. The matching personas could be company employees, data scientists or data intermediaries lacking proper ontologies to make coherent profiles of data they are working on. Re-using structured and detailed schema resources as their knowledge base would save them time and mental effort.

The **fifth** scenario highlights the importance of inspecting already annotated vocabularies to integrate data as much as possible, aiming at data access maximization. Personas are then interested in providing resources that can be translated and preserve semantic and syntactic coherence in the process. Given the heterogeneity problem affecting data and its representations, the possibility of navigating language resources as a single core becomes of the uttermost importance.

The **sixth** scenario accounts for an essential feature of a web portal, the community that supports it with their work. Search information about researchers, their projects and expertise is a key feature of a platform that aims at expanding by increasingly collecting resources and standardizing them according to a precise methodology. Explicit and up-to-date references to

community members' contacts, links to publications and projects make the community live and act as reachable network from an external user, which in this case is the general persona that interacts with the portal to gatehr information about people behind projects.

For each above mentioned scenario, in the attached csv file "PurposeFormalisation", concrete examples and actual personas are given along with their detailed competency questions, which the Data Scientia web portal should be able to answer, when the KGE process is complete and we can query the Metadata KG.

In the building process of scenarios, personas and CQ's, we have thought analytically about the most important actions that users may want to perform when interacting with the web portal and also why they would choose Data Scientia instead of - or along with - other Open Data Catalogs. Real-life use cases and examples proliferated to find key metadata that would give a straightforward access to as much resources as possible offered by Data Scientia. We have then summarized the individual and actual information into what the portal should be able to potentially provide: we have performed a synthesis of all the identified and personalised features to qualify Data Scientia common, core and contextual entities.

We have pinned down four essential common entities: **DataScientia**, the root; **Webpage**, the access point; **Project**, the resource; and **Person**, the one who interacts with the root, getting access to resource.

Core entities represent our core business if we intend Data Scientia as a service: what makes our web portal stand out and adds value is the informational layering: **Live Language** is the container of projects' language layers, **Live Knowledge** of knowledge layers and **Live Data** of data layers. **Live People**, which makes all community members reachable should be a core entity, given the focus on share-ability and re-usability of existing projects. In particular, we have decided to have a **community member** entity to emphasize the openness of the Data Scientia research group, constantly growing and evolving; we have also differentiated community members into the entities **Author** and **Researcher**: Authors are characterized by the fact that they are projects owners and we choose Researcher to describe a DS community member that isn't linked to any project. This distinction is still under scrutiny because authors could be researchers as well, but our priority is now shifted on projects' creators rather than the organisation members. Our last core entity is **Theme**, which tries to subsume the complex concepts of purpose and project description. For now, the web portal doesn't only contain KGE projects, compliant with iTelos, with a specified purpose, but also third-party and bigger DS projects with different descriptions and tags. Theme should ensure the projects and layers chaining according to the common purpose or description of the resources.

We think that the project **Maintainer** should be a contextual entity, because he/she may be addressed in domain-specific contexts, along with a community member **Contact**, which should be taken into account only when a communication channel is opened.

5 Inception

Our inception phase started with the analysis of each entity type highlighted in the previous paragraph and we defined a proto-schema containing essential e-type properties and correlations. We have outlined three e-types: **Webpage**, **Project** and **Person**. We have noted the community relevance of the Data Scientia web portal, a feature that stood out both in the current *Homepage* and *liveschema.eu* catalog of resources.

While designing the web portal structure, from the point of view of a service user, we have realised that an external user does not have to join the community to use some provided services. For this reason, a first distinction was made: a **Person** who uses our service may (or may not) be a **Community Member**. This sub-class has as its signature property *Contact*, a way to be easily reachable. Due to the fact that there are many possible contacts that can be listed for a single person, and considering that not every Person has any possible contact, we decided to transform this property into a e-type.

We have then focused our attention to the e-type **Project**. Our initial idea was to consider three possible Project sub-classes:

- KGE Student's project: the properly layered projects. They should represent the project standard and state of the art: in the future, projects should be uploaded on the Data Scientia following the metadata properties specifically designed for them.
- Data Scientia project: this group includes all the works already present on both *liveschema.eu* and *Homepage*. They are not layered but contain useful information that could be re-used to characterize distinct layers and domain-specific metadata.
- Third Party Project: embedded projects coming from providers outside the community, which do not belong to the previously defined sub-classes.

However, assuming that all **Project** resources should have informational layers in the future, we have decided to consider only the super-class **Project** for now, creating metadata properties about the presence and composition of the three main layers (Data, Knowledge, Language).

Another fundamental aspect of our purpose is the linkage between projects and their author and/or other projects related to the same theme, so we have chosen to add two new e-types:

- **Author**: a **Community Member** extension that has the property *Project*, containing projects associated with the Author under consideration.
- **Theme**: an e-type that enables chaining upon a specific domain of interest. In our formalization, Theme definition allows not only project-to-project links, but also schema-to-schema, language-to-language and data-to-data relationships.

We have decided to add the **Maintainer** e-type. Similarly to **Author**, a Maintainer is an extension of **Community Member**, characterized by *Maintained_Project*, a property which lists

all projects this figure has keep up-to-date and under surveillance. We wanted also to include as a **Community Member** a person directly interacting with Data Scientia which is neither an **Author** nor a **Maintainer**: we have thus created the e-type **Researcher**, a sub-class of **Community Member** with **Publication** as a special property.

Finally, we have worked on the **WebPage** e-type. The latter contains the web interface structure of the whole portal, its sitemap and tree sub-divisions. Through the portal different components and user-friendly webpage interfaces, a general user should be able to browse all projects and inspect the related informational layers. Users should be able to search for **Community Member** instances and obtain their **Contact** or other useful information, such as their e-mail and Linkedin profiles. To achieve this goal, we are developing the following e-types:

- **Live People**: a web page container of community members and eventually their projects and contact information.
- **Live Language**: a web page storage unit for projects' language layers.
- **Live Data**: a web page collector of data resources, such as the actual datasets, pointing at actual data values.
- **Live Knowledge**: a web page catalog, containing all knowledge base resources.

We want to note that Live Language, Live Data and Live Knowledge contain their corresponding layer for each project, and references to the remaining layers. We also wanted to bring to your attention the possible sub-division of the **Knowledge layer**, which makes it an e-type. It can contain both the ontology and teleology layers as properties: this feature depends on how in-depth the ontological analysis for the underlying KG is carried out by the knowledge engineer and the quality of the metadata created.

All steps described so far led to the construction of the Data Scientia portal structure e-types, summarized in the network-like skeleton in Figure 2.

Since we have set the main schema e-types and properties, we now move to data sources, populating all e-type labels with their corresponding entities; we have tried to use real data when possible, and placeholders or fictional hybrids since not all available resources match our e-types definitions.

Protegé and Karma tools helped us produce a reduced example of our KG schema (or Knowledge Layer), associated with the actual datasets of our KG Data Layer.

We are still considering some different schema interpretation that could possibly improve our present structure, such as a new definition of **Author** and/or **Researcher**, to avoid overlays, as well as splitting the existing **Project** *Description* into two property values: *Purpose*, containing the actual purpose of a project, and *Description* that should give a general overview of the work (even if we found that often the distinction between this two properties is not trivial). We are also taking into consideration the introduction of another e-type called **Layer**, which

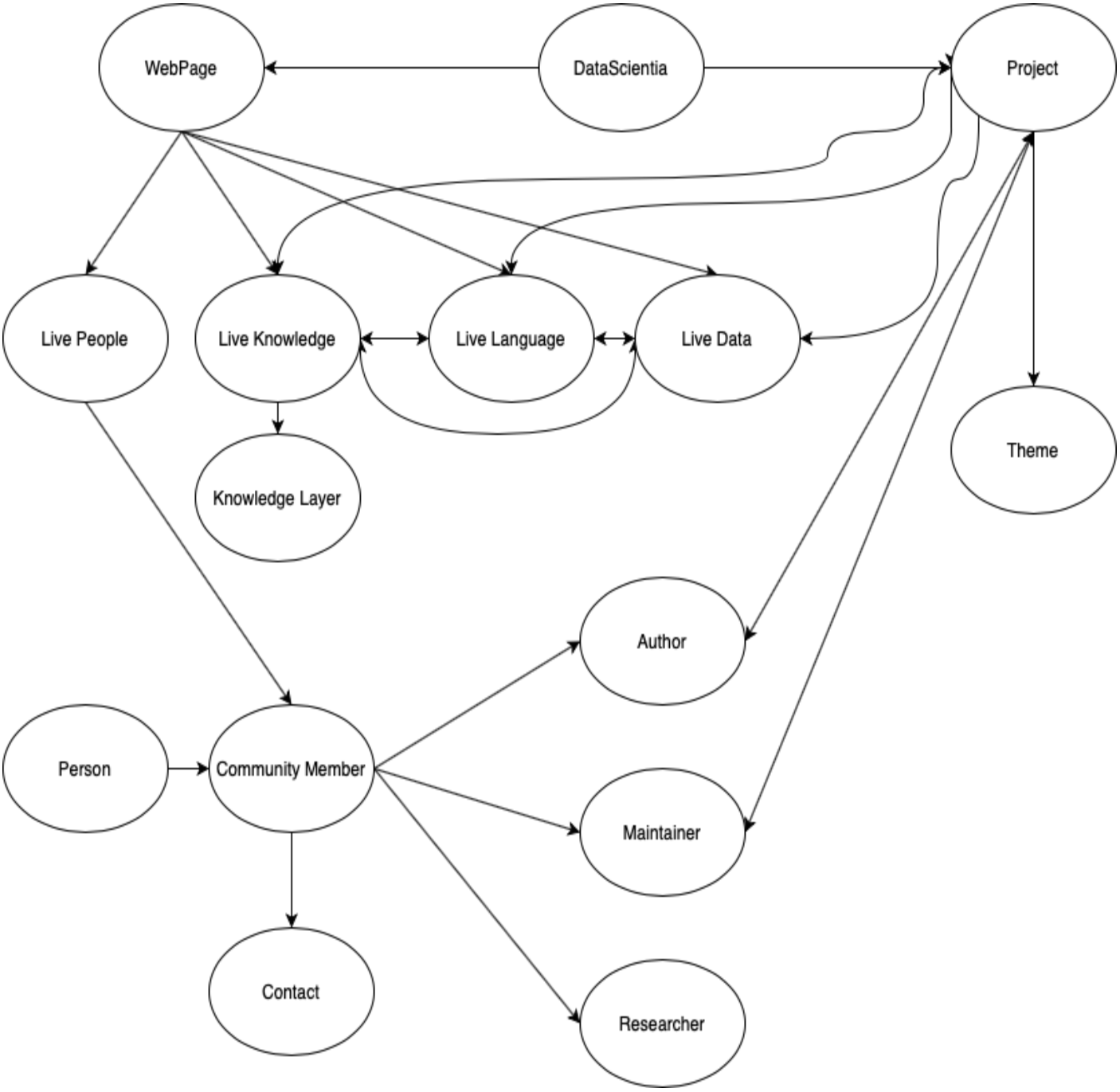


Figure 2: New DataScientia e-type Skeleton

contains the above mentioned informational layers for each project. This action could slim the **Project** table, gaining a further data abstraction level. We are aware of the fact that multiple values in a single table cell may cause some issues: for example, an **Author** of many projects or a **Project** with several authors. We are working to solve the eventual collisions, if they arise. Due to this problem and also the absence of actual datasets to look into, we have not yet exploited Protegé formatting and integration tools.

References

[Giu+21] Fausto Giunchiglia et al. *iTelos – Purpose Driven Knowledge Graph Generation*. 2021. doi: 10.48550/ARXIV.2105.09418. url: https://arxiv.org/abs/2105.09418.