Towards Purposeful Reuse of Semantic Datasets Through Goal-Driven Summarization

Panos Alexopoulos and José-Manuel Gómez Pérez
iSOCO, Intelligent Software Components S.A.
Madrid, Spain,
e-mail: {palexopoulos, jmgomez}@isoco.com

***Abstract*—The emergence in the last years of initiatives like the Linked Open Data (LOD) has led to a significant increase of the amount of structured semantic data on the Web. Nevertheless, the wider reuse of such public semantic data is inhibited by the difficulty for users to decide whether a given dataset is actually suitable for their needs. This is because semantic datasets typically cover diverse domains, do not follow a unified way of organizing the knowledge and may differ in a number of dimensions. With that in mind, in this paper, we report our work in progress on a goal-driven dataset summarization approach that may facilitate better understanding and reuse-oriented evaluation of available semantic data.**

Keywords-Semantic Data Reuse; Semantic Data Summarization.

I. INTRODUCTION

The emergence in the last years of initiatives like the Linked Open Data (LOD) [1] has led to a significant increase of the amount of structured semantic datasets on the Web. Nevertheless, while this increased availability of such datasets yields various opportunities for organizations and technical professionals to derive added value from them, their wide heterogeneity and underlying complexity makes their practical use and exploitation quite difficult and challenging. For that, solutions that can enable the better understanding and easier consumption of semantic datasets are of crucial importance.

The typical use case scenario we consider in this paper assumes some organization that wants to reuse public semantic datasets to i) enrich with them its own data so as to make the latter more usable and increase its usability and value and ii) utilize the enriched data within knowledge intensive applications for particular purposes (e.g., decision support). Such tasks are typically performed by knowledge engineers and the common problem associated to them is the so called **knowledge acquisition bottleneck**, namely, the high amount of time and effort required to acquire and maintain the needed knowledge [2].

Our position is that the reuse of existing public semantic data can be a promising way to (partially) alleviate the knowledge acquisition problem. One reason for that is that the volume and diversity of public semantic datasets are increasing at high rates [1], resulting into a large amount of both generic and domain-specific knowledge that is available to use for various application scenarios. Another advantage of the reuse approach is that the maintenance and evolution of these datasets is the responsibility of their publishers, thus reducing the required efforts and costs for this task in the organization's side.

As an example of this, consider a sport news organization that wants to create and maintain a knowledge base about the Spanish football league (teams, rosters, results, etc.). The pace at which this knowledge changes is quite fast (e.g., team rosters change at least every year, sometimes even more frequently), meaning that the organization needs to have a dedicated team that constantly monitors these changes and updates the knowledge base. As much of this information is already available in public semantic datasets and, more importantly, it is (almost) always up to date, it would be better for the organization to reuse this data instead of creating it from scratch and having to maintain it.

Nevertheless, an important problem that inhibits the wider reuse of such public semantic data is the difficulty for knowledge engineers to decide whether a given dataset is actually suitable for their needs. This is because semantic datasets typically cover diverse domains, do not follow a unified way of organizing the knowledge and differ in a number of features including size, coverage, granularity and descriptiveness. This makes the task of assessing whether a dataset satisfies particular requirements (e.g., covering adequately a particular domain) and/or comparing different datasets to select which one is more suitable for a given purpose quite difficult.

For instance, in the example mentioned above about data related to the Spanish football league, one may find such data in DBPedia[12] and Freebase[11]. To evaluate these sources, the knowledge engineer needs to examine and assess a variety of factors including i) the domain's coverage, namely, the degree to which the containing data cover the Spanish football league (e.g., one of the sources might not contain adequate data for a given year), or ii) the dataset's consistency, namely, the absence of contradictions in the data (e.g., there might be statements suggesting that a player is currently playing for two clubs).

As a way to tackle this problem, we envision the development of a framework that will enable users to derive **semantic data summaries**, namely useful descriptions, measures

 ISBN: 978-1-61208-293-6

and indicators that provide a landscape yet informative view on a dataset that enables the assessment of the latter's potential value. This task of semantic data summarization is rather overlooked in the research community and has only been addressed by a few works, e.g., [3] [4] [5], each of which generates dataset summaries according to different data features and by applying different criteria.

Yet, the problem with these approaches is that they treat the summarization task in an application and user independent way by producing generic summaries whose usefulness is limited to an all-purpose very high level overview of the data. By contrast, in our scenario, we are interested in facilitating the generation of requirements-oriented and task-specific summaries that may be significantly more helpful to the knowledge engineers and data practitioners in their task to locate semantic data to reuse and exploit.

To that end, in this paper, we report our work in progress on a goal-driven data summarization framework that may be used to examine and evaluate the suitability of semantic data sources for reuse in particular application domains and scenarios. Within this framework users are able to define and execute custom summarization processes to generate useful dataset summaries. A custom summarization process can be seen as an orchestration of primitive predefined parameterizable data analysis processes each of which may deal with a different aspect of the data. More importantly, such a process is linked to a particular goal/problem/need that it is supposed to serve, thus forming a reusable knowledge component that can be shared among multiple users with similar needs.

The structure of the rest paper is as follows: In the next section, we outline the key aspects of our approach and the basic components of our summarization framework. In Section III, we discuss a particular small-scale application of our framework in a dataset evaluation scenario, and, in Section V, we conclude and outline our future work plans.

II. Semantic Dataset Summarization Framework

Our proposed summarization framework aims to enable its intended users to answer the following question: *"Given an application scenario where semantic data is required, how suitable is a given existing dataset for the purposes of this scenario?"*. To answer this question, users normally need to be able to: i) explicitly express the requirements that a dataset needs to satisfy for a given task or goal and ii) automatically measure/assess the extent to which a dataset satisfies each of these requirements and compile a summary report.

To implement these two capabilities, we follow a *checklist-based* approach. Checklists are practically lists of action items arranged in a systematic manner that allow users to record the completion of each of them and they are widely applied across multiple industries, like healthcare or aviation, to ensure reliable and consistent execution of complex operations [6]. In our case, we apply checklists to define and execute custom dataset summarization tasks in the form of lists of goal-specific requirements and associated summarization processes. In the following paragraphs, we explain how such tasks and processes may be represented, created and used.

A. Summarization Task Representation

To represent custom summarization tasks according to the aforementioned checklist paradigm, we adopt the *Minim model* [7] that allows us to represent for concrete instances of summarization tasks the following information:

- The **Goals** the dataset summarization task is designed to serve. In the Minim's terminology [7], these are called constraints and they are used to denote the purpose of the summarization task and the intended use of the produced summary. This is important as different tasks may have different purposes (e.g., the requirements for checking whether a dataset is appropriate for disambiguation may be different from those required for question answering) and, thus, the goal-related information is crucial for selecting an already defined task in a given application scenario.
- The **Requirements** (or checklist entries) against which the summarization task evaluates the dataset. For example, we may wish to assess whether a dataset contains particular information about a given domain or topic or that it satisfies particular quality criteria (e.g., consistency). The number and nature of the requirements depend practically on the goal of the summarization task and thus they may be substantially different among different application scenarios.
- The **Data Analysis Operations** that the summarization task employs in order to assess the satisfaction of its requirements. In the Minim's terminology, these operations are called rules and practically they take many forms, from simple execution of queries to complex data processing and analysis algorithms like graph analysis or topic modeling. The assessment of a given requirement may require the execution of multiple operations while the same operation may be used to assess multiple requirements.

B. Summarization Task Creation

To create a summarization task one needs to define its goal(s), its requirements and the associated to these operations. Some high-level requirements that we have already identified and they may be used for multiple goals are the following:

- **Evaluate the dataset's coverage of a particular domain/topic**: This requirement aims to measure the extent to which a dataset describes a given domain or topic. This can be at schema level (e.g., how many and which concepts or relations are defined), at instance level (e.g., how many and which instances of a given

 ISBN: 978-1-61208-293-6

concept or relation does the dataset have) or with more complex operations (e.g., comparison with a corpus).

- **Evaluate the dataset's labeling adequacy and richness**: This requirement aims to measure the extent to which the dataset's elements (concepts, instances, relations etc.) are accompanied by representative and comprehensible labels, in one or more languages. This can be useful to assess two things: i) the comprehensibility of the data, i.e., the ease with which human consumers can understand and utilize the data and ii) the quality and usefulness of a dataset as a term thesaurus.
- **Evaluate Connectivity**: This requirement checks the existence of paths between concepts or entities, i.e., whether it is possible to go from a given concept to another on the graph and in what ways. This is can be an important aspect of a dataset related, for example, to its ability to answer queries involving particular related entities.

Each of the above requirements can be implemented by means of one or more data analysis operations. Some operations we have already defined for our framework are the following:

- Check the existence of a particular element (concept, relation, attribute, instance, axiom) in the dataset or of a relational path between particular concepts or instances.
- Measure the number of ambiguous entities in the dataset.
- Measure the number of labeled entities.

C. Dataset Summary Generation

For the generation of goal-specific dataset summaries, we are currently developing a tool that may take as input one or more datasets and a summary goal and run on them specified summarization tasks that correspond to this goal. The output of this tool should be a detailed report about the input datasets, describing whether and to what extend do they satisfy each requirement. The next section provides a concrete example of this output in the context of an actual use case where we applied our framework.

III. Framework Application

A concrete scenario where we applied our framework involved the assessment of public datasets for the purposes of reusing them within a semantic annotation system. In particular, we wanted to annotate texts describing football matches from the Spanish League by means of an in-house ontology-based semantic entity recognition system whose effectiveness depends on the characteristics and quality of the available domain knowledge. For that, we wanted the dataset to be reused to i) contain information about all the current teams of the Spanish football league, ii) all its entities to have at least one associated label and iii) to relate teams with the players that current play in them.

```
<#contains_all_spanish_liga_teams> a minim:Requirement
<#has_labels_for_all_entities> a minim:Requirement
<#has_player_team_relation> a minim:Requirement

<#summarization_task_for_text_annotation> a minim:Model
  rdfs:label "Summarization task for text annotation"
  <minim:hasMustRequirement
    <#contains_all_spanish_liga_teams>,
    <#has_labels_for_all_entities>,
    <#has_player_team_relation>

<#used_for_text_annotation> a minim:Constraint
  minim:forPurpose "Text Annotation"
  minim:toModel
   <#summarization_task_for_text_annotation>
```

Figure 1. Example of Formal Summarization Task Definition

To perform this assessment, we used the model of section II to define a custom summarization task that could help us assess the degree to which some existing datasets satisfied these requirements. A snapshot of the formal definition of the task where the task, its goal and its requirements are defined, is shown in Figure 1.

We executed this task against DBPedia and Freebase, automatically producing the summary report of table I. As one can see the system provides a yes/no answer as to whether each dataset satisfies each requirement but also additional information on why this may or may not be the case (e.g., the percentage of missing labels). The first reason this latter feature is important is that a requirement might not be satisfied because the relevant threshold might have been set too high (e.g., the requirement for 100% labeling). Thus, by showing the actual satisfaction score, the user may decide to relax his/her constraints for the given requirement, especially when there is no dataset fully satisfying it. The second reason is that a requirement might seem to be satisfied, yet that might not be actually true for reasons pertaining to the system's underlying methods and/or the datasets. For example, a closer inspection of the current roster relation in Freebase's website reveals that its instances do not adhere to the semantics of the relation as there are player-team pairs that are no longer valid. Thus, the generated summaries allow users to judge further the suitability of the datasets and refine the requirement rules.

IV. Related Work

Most approaches for semantic data summarization focus on deriving generic goal-independent summaries that provide a high level overview of the data and highlight some of its aspects. For instance, in [3], summaries have the form of questions that can be answered by the dataset, while in [4] summaries consist of the most representative concepts of an ontology, determined based on cognitive and statistical criteria. Nevertheless, these types of summaries are not linked to particular goals nor are they parameterizable.

Relevant to ours work may be also found in the area of semantic data quality where various approaches attempt to define quality criteria and metrics for semantic data. SemRef

 ISBN: 978-1-61208-293-6

Table I
EXAMPLE OF A GOAL-DRIVEN DATASET SUMMARY

Requirement	**DBPedia**	**Freebase**
Spanish League Coverage	YES	YES
At least one label per entity	NO (5% of the entities has no labels)	YES
Player-Team Relation	YES (*"dbpprop:currentclub"*)	YES (*"http://freebase.com/soccer/football_team/current_roster", "http://freebase.com/soccer/football_player/current_team"*)

[8], for example, defines such criteria for evaluating the quality of semantic metadata with respect to how well they describe a set of resources. A more generic framework is *Sieve* [9] that allows the definition and calculation of custom quality metrics over already available dataset metadata. In that sense it is similar to our approach as it is parameterizable and goal-driven. Nevertheless, our framework goes one step further by allowing also the definition of generation methods for this metadata (in the form of the data analysis operations), thus covering a wider set of use cases.

Finally, checklist-based approaches have been recently used in biology [10] and in scientific workflows [7], though not yet, to the best of our knowledge, for the task of summarizing and evaluating semantic datasets for reuse purposes.

V. Conclusion and Future Work

In this paper, we presented our ongoing work on a framework for the definition and execution of goal-driven semantic data summarization tasks, as a way to enable organizations and practitioners to take better decisions on whether existing datasets are suitable for their purposes. The framework follows the checklist paradigm and uses a formal ontological model to represent summarization tasks by means of goals, requirements and data analysis operations. Our immediate future works include further technical development of the framework, especially in relation to the management of the datasets (a list of available datasets needs to be created and maintained from sites like http://linkeddata.org/datasets, while local endpoints should be created for datasets that currently lack ones). Moreover, additional high-level requirements and data analysis operations will be defined, as well as a User Interface for the definition and generation of semantic data summaries.

Acknowledgment

The research leading to this results has received funding from the People Programme (Marie Curie Actions) of the European Union's 7th Framework Programme P7/2007-2013 under REA grant agreement n^o 286348.

References

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," International Journal of Semantic Web Information Systems, vol. 5, no. 3, 2009, pp. 1-22.

[2] S. Szumlanski and F. Gomez, "Automatically acquiring a semantic network of related concepts," in Proceedings of the 19th ACM international conference on Information and knowledge management. New York, NY, USA, ACM, 2010, pp. 19-28.

[3] M. d'Aquin and E. Motta, "Extracting relevant questions to an RDF dataset using formal concept analysis." in K-CAP, M. A. Musen and O. Corcho, Eds. ACM, 2011, pp. 121-128.

[4] S. Peroni, E. Motta, and M. dAquin, "Identifying key concepts in an ontology through the integration of cognitive principles with statistical and topological measures," in Third Asian Semantic Web Conference, Bangkok, Thailand, 2008.

[5] V. Presutti, L. Aroyo, A. Adamou, B. Schopman, A. Gangemi, and G. Schreiber, "Extracting core knowledge from linked data," in Proceedings of the Second Workshop on Consuming Linked Data, COLD2011, Workshop in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011). CEUR-WS.

[6] B. Hales and P. Pronovost, "The checklista tool for error management and performance improvement," Journal of Critical Care, vol. 21, no. 3, Sep. 2006, pp. 231-235.

[7] K. Belhajjame, M. Roos, E. Garcia-Cuesta, G. Klyne, J. Zhao, D. De Roure, C. Goble, J. M. Gomez-Perez, K. Hettne, A. Garrido, "Why workflows break - understanding and combating decay in taverna workflows." in Proceedings of the 2012 IEEE 8th International Conference on E-Science. IEEE Computer Society, 2012, pp. 1-9.

[8] Y. Lei, V. Uren, and E. Motta, "A framework for evaluating semantic metadata," in Proceedings of the 4th international conference on Knowledge Capture, K-CAP 2007, 2007, pp. 135-142.

[9] P. N. Mendes, H. Muhleisen, and C. Bizer, "Sieve: Linked Data Quality Assessment and Fusion," in 2nd International Workshop on Linked Web Data Management (LWDM 2012) at the 15th International Conference on Extending Database Technology, EDBT 2012, March.

[10] C.F. Taylor, N.W. Paton , K. S. Lilley, P. A. Binz, R. K. Julian Jr, A.R., Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M-. J. Dunn, A. J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S.D. Patterson, P. Ping, S.L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M- R. Wilkins, I. Xenarios, J.R. Yates 3rd, H. Hermjakob., "The minimum information about a proteomics experiment (MIAPE)." Nature biotechnology, no. 8, Aug. 2007, pp. 887-893.

[11] Freebase, http://www.freebase.com,Accessed: 16/09/2013.

[12] DBPedia, http://dbpedia.org, Accessed: 16/09/2013.

 ISBN: 978-1-61208-293-6