Representing Texts as Contextualized Entity-Centric Linked Data Graphs

André Freitas, Seán O'Riain, Edward Curry
Digital Enterprise Researh Institute (DERI)
National University of Ireland, Galway

João C. P. da Silva, Danilo S. Carvalho
Computer Science Department
Federal University of Rio de Janeiro (UFRJ)

***Abstract*—The integration of a small fraction of the information present in the Web of Documents to the Linked Data Web can provide a significant shift on the amount of information available to data consumers. However, information extracted from text does not easily fit into the usually highly normalized structure of ontology-based datasets. While the representation of structured data assumes a high level of regularity, relatively simple and consistent conceptual models, the representation of information extracted from texts need to take into account large terminological variation, complex contextual/dependency patterns, and fuzzy or conflicting semantics. This work focuses on bridging the gap between structured and unstructured data, proposing the representation of text as structured discourse graphs (SDGs), targeting an RDF representation of unstructured data. The representation focuses on a semantic best-effort information extraction scenario, where information from text is extracted under a pay-as-you-go data quality perspective, trading terminological normalization for domain-independency, context capture, wider representation scope and maximization of textual information capture.**

I. Introduction

The Linked Data Web brings the vision of a Web-scale semantic data graph layer which can improve the ability of users and systems to access and semantically interpret information. Most of the information available on the Web today is in an unstructured text format. The integration of this information into the Linked Data Web is a fundamental step towards enabling the Semantic Web vision. The semantics of unstructured text, however, does not easily fit into structured datasets. While the representation of structured data assumes a high level of regularity and normalization, relatively simple conceptual models and a consensual semantics between the users of a structured dataset, the representation of information extracted from texts need to take into account large terminological variation, complex context patterns, fuzzy and conflicting semantics and intrinsically ambiguous sentences.

Most information extraction (IE) approaches targeting the extraction of relations from unstructured text have either focused on the extraction of binary relations (triples) or on specific relation patterns which are going to feed a well-structured ontology (e.g. events), scenarios where accuracy, consistency and a high level of lexical and structural normalization are primary concerns. These IE approaches can be complemented by alternative information extraction scenarios where accuracy, consistency and regularity are traded by domain-independency, context capture, wider extraction scope and maximization of the text semantics representation, where data semantics and data quality are built and improved over time (under a *pay-as-you-go* data quality perspective). This type of approach can provide a complementary semantic layer to the Web, enriching existing datasets, bridging the gap between the Linked Data Web and the Web of Documents. From a representation perspective, many issues arise when information extracted from texts is aimed to be represented on the Linked Data/Semantic Web context.

This work aims at providing a *structured discourse graph* (SDG) model which can be used to improve the semantic integration, representation and interpretation of unstructured texts within the context of the Linked Data Web. Despite the availability of text representation models from computational linguistics such as Discourse Representation Structures (DRS) there is still a major gap regarding the representation of discourse elements under a data model perspective. The core features of the proposed representation includes (i) an *entity-centric data representation model*, which facilitates the integration and alignment of discourse elements with entities on the Linked Data Web (ii) an ontology/vocabulary agnostic representation, where there is no commitment to a specific ontological/conceptual model, (iii) a flexible contextual representation in natural language, (iv) the formulation of an algorithmic interpretation model for SDGs based on the concept of graph traversal and (v) a discussion on the representation of SDGs as RDF(S) and Linked Data. The proposed model focuses on the representation of *complex factual statements* in which mappings to triples is non-trivial.

An additional goal of this work is to contextualize this discussion under a semantic best-effort (SBE) information extraction [3] angle. This concern is motivated by the perspective that, in practice, the complexity of open information extraction task demands a representation model robust to potential extraction errors or to information incompleteness. Additionally the representation should support the semantic refinement and evolution under a pay-as-you-go data quality and data integration perspective. These goals demand the formulation of a principled semantic representation which should accommodate these requirements.

II. Contextualized Entity-Centric Graphs: Representation Requirements

To achieve a representation which provides the previously described capabilities, a set of requirements are defined for the structured discourse graph (SDG) representation.

1. Entity-centric graph model: An entity pivot is a *named entity* present in the subject or object part of a statement. The

isolation of entity pivots into specific graph elements allows (i) the creation of an *entity-centric data model*, where the information, initially centered on documents becomes centered on entities and (ii) a *semantic interpretation approach*, where the less ambiguous part of the semantic interpretation process is prioritized using the entity as a semantic pivot.

2. Principled & maximized representation of text semantics: The transference between the information present in a sentence to the graph representation should be maximized. Lexical normalization of predicates and classes is optional. The removal of a terminological normalization constraint allows the maximization of the text extraction. The graph representation should also support an algorithmic interpretation of the extracted SDG.

3. Maximization of the correctness of the syntactic-structural mapping: Syntactic structures of the parsed text should correctly map to its corresponding graph structure.

4. Conceptual model independency: The extracted graphs shoud not commit to a specific ontology/vocabulary model. The use of conceptual models, reduces the generality of the extraction representation.

5. Context capture & representation: Contextual information related to a triple statement should be represented in the extracted graph, to allow a contextualized semantic interpretation. Contextual statements (such as temporality) define the context in which another statement holds. Dependencies between sentences in the text should also be made explicit in the final representation. Context can also be defined by semantic dependencies which could be intra or inter-sentence.

6. Pay-as-you-go semantic reference: Unstructured text may contain complex semantic dependencies. The extraction graph should support an extensible representation of semantic and co-referential dependencies and should support the evolution and refinement of the semantic model (pay-as-you-go dependency resolution).

7. Standardized representation compatibility: The representation should maximize its compatibility with a standards-based data representation format to facilitate the graph integration and interoperability on the Web.

III. Related Work

Different works targeted the representation of relations extracted from texts. Discourse Representation Structure (DRS) is a semantic structured language for the representation of natural language sentences. Presutti et al. [5] propose a method for mapping DRS into RDF/OWL for ontology learning and population (OL&P), and introduce FRED, a tool for performing OL&P over texts. Open information extraction approaches [2] have concentrated on the extraction of single relations or specific patterns which are mapped to ontology and vocabularies. Harrington & Wojtinnek [1] propose SemML, an XML serialization format for semantic networks. The main motivation for the creation of SemML is to provide a serialization format which better supports nested contextual elements such as temporal and triple annotations (e.g. associated numeric values to triples). NLP Interchange Format (NIF)[1] is a format which targets the interoperability among NLP tools, language resources and annotations. NIF consists of two vocabularies (the String Ontology and the Structured Sentence Ontology) which allows the annotation of documents and sentences. NIF concentrates on the documentation of the workflow of resources which are used on the analysis.

[1]http://nlp2rdf.org/nif-1-0

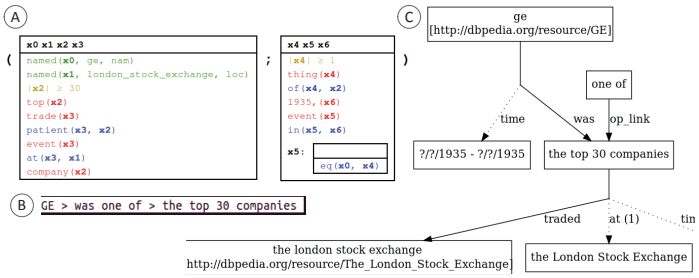

Fig. 1. Comparison between different representation models: (A) discourse representation structure, (B) single triples relation extraction (ReVerb), (C) structured discourse graphs (SDGs).

Comparatively, this work focuses on providing a principled description on a representation model complementary to existing approaches, focusing on a graph representation approach for texts which can be easily integrated to the Linked Data Web (entity-centric), which is able to capture and represent complex contextual structures, with no commitment to a specific conceptual model and focused on a semantic best-effort scenario. Figure 1 provides a comparison among different representation and extraction perspectives.

IV. A Semantic Model for Representing Unstructured Text as Graphs

The representation of text as a structured discourse graph (SDG) following the previous requirements, demands the definition of a principled semantic model for SDGs. This section analyzes and defines the mapping between grammatical roles, syntactic structures and text structures to a SDG semantic model. Section IV.A analyses and describes the elements of the semantic model, which are formalized in section IV.B.

A. Semantic Model Elements

Named Entities: Named entities refer to the description of entities for which one or many rigid designators stands for the referent. Rigid designators include categories such as proper nouns, temporal expressions, biological species, substances, etc. A named entity is defined by one or more proper nouns (**NNP**) in a noun phrase (**NP**). In RDF named entities map to instances. All graphs in Figure 2(A)-(I) contain named entities. Impacts requirement: 1.

Non-named (generic) entities: Non-named entities map to *non-rigid designators*. Non-named entities (e.g. 'President of the United States') are more subject to vocabulary variation, i.e. polysemy and homonymy. Additionally, non-named entities have more complex compositional patterns: commonly non-named entities are composed with less specific named or non-named entities, which can be referenced in different contexts. A non-named entity is defined by one or more nouns (**NN**),

adjectives(**JJ**) in a noun phrase (**NP**). In RDF a non-named entity maps to classes. Ex.: Node '13th District' in Figure 2(D). Impacts requirements: 1,3.

Properties: Properties are built from verbs (**VB**) or from passive verb constructions (e.g. is supported by). Graph Pattern: Figure 3(1). Ex.: Edges from all graphs in Figure 2. Impacts requirements: 2,3,4.

Quantifiers & Generic Operators: Represent a special category of nodes which provide an additional qualification over named or non-named entities. Both quantifiers and generic operators are specified by an enumerated set of elements (from named, non-named entities and properties which maps to an open set of terms), which maps to *adverbs*, *numbers*, *comparative and superlative* (suffixes and modifiers). Examples of quantifiers are: *Quantifier:* e.g. one, two, (cardinal numbers), many (much), some, all, thousands of, one of, several, only, most of; *Negation:* e.g. not *Modal:* e.g. could, may, shall, need to, have to, must, maybe, always, possibly; *Comparative:* e.g. largest, smallest, most, largest, smallest, the same, is equal, like, similar to, more than, less than. The graph pattern in Figure 3(5) shows the core structure of a triple with an operator. Ex.: Figure 2(E). Impacts requirements: 2,3,5,6.

Triple Trees: Not all facts extracted from a sentence can be represented in one triple. On a normalized dataset scenario, one semantic statement which demands more than one triple is mapped to a concept model structure (as in the case of events for example) which is not explicitly present in the discourse. In the SDG representation, sentences which demands more than one triple can be organized into a triple tree. A triple tree is built by a transformation from the syntactic tree of a sentence to a set of triples, where the sentence subject defines the root node of the triple tree. The interpretation of a triple tree is defined by a complete DFS traversal of the tree following the interpretation patterns in section IV.B), where each connected path from the root node to a non-root node defines an *interpretation path*. Graph Pattern: Figure 3(1). Ex.: Figure 2(C). Impacts requirements: 2,3,4.

Context elements: A fact extracted from a natural language text demands a semantic interpretation which may depend on different contexts where the fact is embedded (such as a temporal context). In a factual corpus the main contextual information is intra-sentence (given by a different clause in the same sentence). Intra-sentence context for a triple can be represented by the use of reification (Figure 2). Contexts can also be important to define the semantics of an entity present in two or more triple trees. For example, the interpretation of an entity which is neither a root and a leaf node (Figure 3(4)) demands the capture of the pairwise combination of its backwards and forward properties in multiple contexts. This is lost in an uncontextualized graph interpretation process. A third level of context can be defined by mapping the dependencies between extracted triple trees, taking into account the sentences ordering and the relation to text elements in the original discourse. Graph Pattern: Figure 3(4). Ex.: Temporal nodes in Figure 2. Impacts requirements: 5.

Co-Referential elements: Some discourse elements contain indirect references to named entities (co-references). Two types co-references exist: pronominal (where pronouns represent the proxy for a named entity, (e.g. *He* referring to Barack Obama)

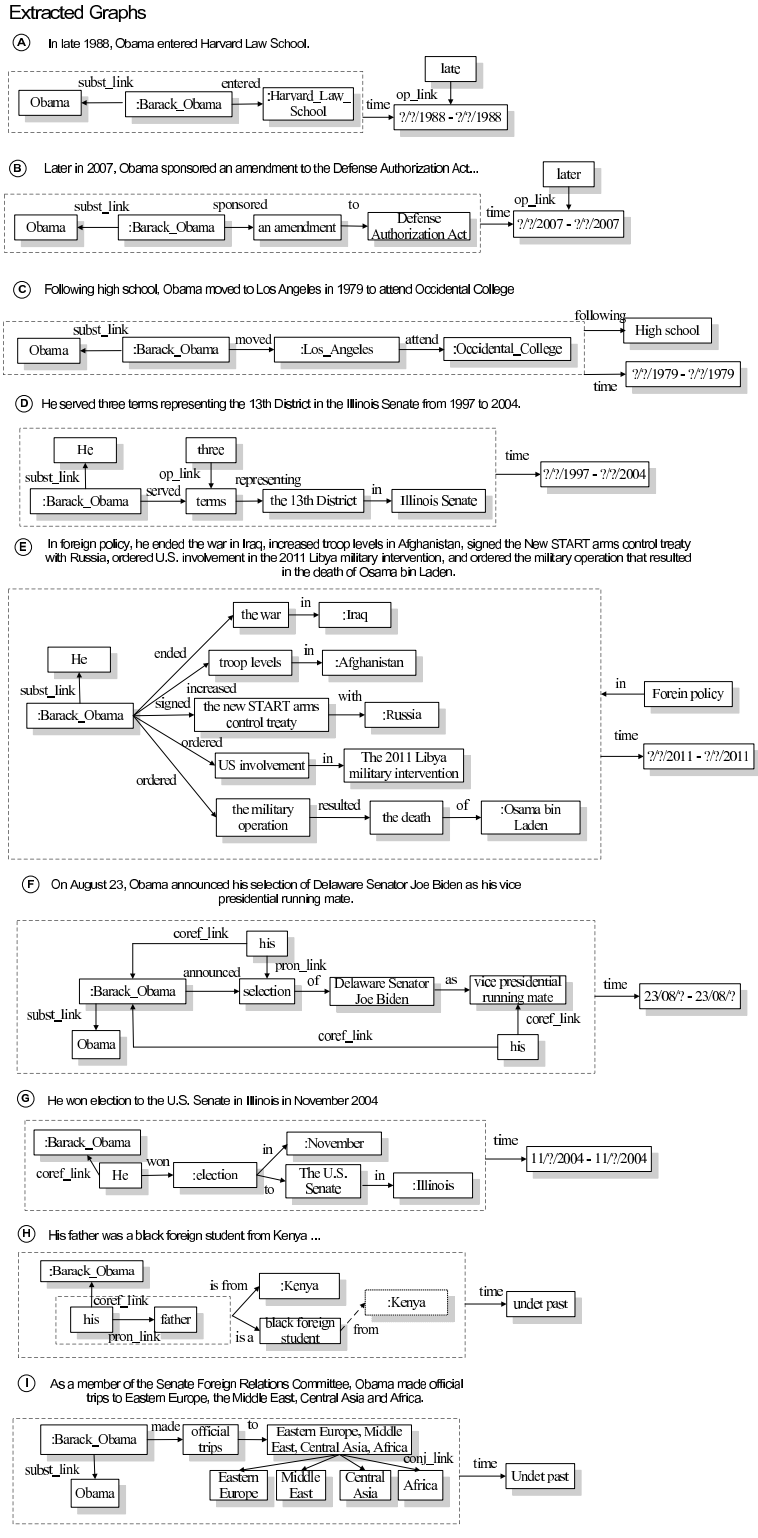

Fig. 2. Examples of extracted sentence graphs from the Wikipedia article *Barack Obama*.

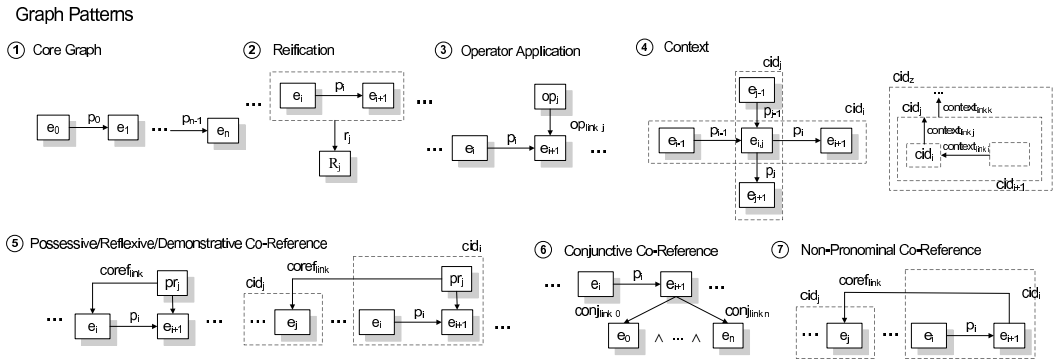

Fig. 3. Depiction of semantic model elements and the graph interpretation patterns.

and non-pronominal, where non-named entities are the proxy (e.g. the 44th President of the United States). Co-references can refer to either intra or inter sentences named entities. While in some cases co-references can be handled by substituting the co-referent term by the named entity (as in personal pronouns), in other cases this direct substitution can corrupt the semantics of the representation (as in the case of reflexive and personal pronouns) or can mask errors in a semantic best-effort extraction scenario. Co-reference terms include: you, I, himself, her, this, that, etc. Graph Pattern: Figure 3(5)(6). Ex.: Figure 2(F)(H)(I). Impacts requirements: 3,5,6.

Resolved & normalized entities: Resolved entities are entities where a node-substitution in the graph was made from a co-reference to a named entity (e.g. a *personal pronoun* to a named entity). Normalized entities are entities which were transformed to a normalized form. A temporal normalization where date & time references are mapped to a standardized format (September 1st of 2010 to 01/09/2010) is an example of temporal normalization. The substitution can be made explicit on the graph. Ex.: Figure 2(A)-(G). Impacts requirements: 1.

B. Definitions

The definitions below formalize the semantic model elements and provide graph patterns and their interpretation for the SDG. A *graph pattern* is an atomic graph structure which maps to a recurrent discourse structure. An *interpretation* consists in a graph traversal sequence of one or more *graph patterns* in a well-defined order. Examples of extracted graphs and the abstract graph patterns are depicted in Figure 2 and Figure 3. Some of the definitions were omitted due to space constraints. An expanded version of the definitions can be found in [6].

Named and Non-named Entities: are represented respectively by ne and $\sim ne$. We use e to denote indistinguishably named/non-named entities. They are interpreted as finite subsets of an infinite set U of IRIs. Thus, $[[ne]] \in U$, $[[\sim ne]] \in U$ and $[[e]] \in U$.

Basic Triple: a triple $tr = (e_s, p, e_o)$ where e_s, e_o represent entities associated respectively with the subject (s) and object (o), and p represents a relation between e_s and e_o. A basic triple is called **core triple** (denoted by tr_c) when $e_s = ne_s$ and $e_o = ne_o$ are both named entitites. Otherwise, it is called **semi-core triple** and denoted by tr_{sc}. The interpretation of a basic triple is such that $[[tr]] = ([[e_s]], [[p]], [[e_o]]) \in U \times U \times U$.

Reification Triple: a triple $tr_{rei} = (tr, rei_{link}, rei_{obj})$ where tr represents a basic triple, rei_{link} represents relation and rei_{obj} represents a *reification object* (i.e., an entity, a value or a triple). A **temporal reification** is a special kind of reification triple where rei_{link} has a special stamp ($time$) and rei_{obj} represents explicit or implicit data references. The interpretation of a reification triple $[[tr_{rei}]] = ([[tr]], [[rei_{link}]], [[rei_{obj}]]) \in U^3 \times U \times U$ means that the basic triple tr is reified in rei_{obj} through relation rei_{link}.

Quantifier Operators & Generic Operators: a triple $opt = (e_o, op_{link}, op)$ where e_o is the object element in a basic triple tr, op represents a specific operator of e_o wrt op_{link}. The interpretation of a quantifier or generic operator is $[[opt]] = ([[proj_3(tr)]], [[op_{link}]], [[op]]) = ([[e_o]], [[op_{link}]], [[op]]) \in U \times U \times U$, where $proj_3(tr)$ is a projection map that takes an element (e_s, p, e_o) to the value e_o, meaning that the quantifier or generic operator op is applied to object e_o through link op_{link}.

Conjunctive Co-Reference: set of triples $ccr = \bigcup_{i=0}^{n}\{(e, conj_{link_i}, ne_i)\}$ which means that the entity e is composed by the conjunction of named entities ne_i. The interpretation of a conjunctive co-reference is $[[ccr]] = \{([[e]], [[conj_{link_i}]], [[ne_i]]) \in U \times U \times U : [[e]] = [[proj_3(tr)]] and \bigwedge_{i=0}^{n}[[ne_i]] sameas\ [[e]]\}$, meaning that the entity e is related through p to e_{i+1} which is formed by the conjunction of ($conj_{link}$) named entities $ne_0, ne_1, \cdots, ne_n$.

Possessive/Reflexive/Demonstrative Co-Reference: set of triples $pcr = \{(\sim ne_i, coref_{link}, pr), (pr, coref_{link}, e_j)\}$ where $coref_{link}$ associates non-named entities $\sim ne_i$ with e_j through the pronoun pr if there is a basic triple $tr = (\sim ne_i, p, e_j)$. The interpretation of this kind of co-reference is $[[pcr]] = \{([[proj_1(tr)]], [[coref_{link}]], [[pr]]),$ $([[pr]], [[coref_{link}]], [[proj_3(tr)]]) \in U \times U \times U : tr = (\sim ne_i, p, e_j)\}$

With these elements we can define an **extracted graph** G from a given corpus as a set of (*basic* and *reified*) *triples* and (*generic*, *quantifier* and *co-reference*) *operators*. In an extracted graph:

basic path P_b: is a sequence of basic triples $P_b =< tr_1, tr_2, \cdots, tr_n >$ such that $tr_i = (e_i, rel_{link_i}, e_{i+1})$ for all $i \in [1, n]$. The interpretation of a basic path $[[P_b]] =< [[e_0]], [[rel_{link_0}]], \cdots, [[rel_{link_{n-1}}]], [[e_n]] >$ is such that for

tr_i and tr_{i+1} $\forall i \in [1, n-1]$, we have $[[proj_3(tr_i)]] = [[proj_1(tr_{i+1})]]$.

reified path P_r: is a basic path such that there are reified triples associated with some of tr_i's in the sequence $P_b =< tr_1, tr_2, \cdots, tr_n >$. The interpretation of a reified path is such that for a basic triple tr_i and reified triples tr_{rei_j} and $tr_{rei_{j+1}}$ and their respective interpretations [6].

operational path P_o: is a basic path such that there are operators associated with some of entities e_i's in triples tr_i's in the sequence $P_b =< tr_1, tr_2, \cdots, tr_n >$. The interpretation of a operational path is such that for a basic or reified triples tr_i and $tr_{i+1} \in P_o$ with their respective interpretations[6].
complex path P_c: contains both reified and operational paths.

The interpretation of a basic triple with operators should be done before the reification when it is also a reified triple. We can associate a context to extracted graphs introducing:

Context Triple: a triple $context = (tr, context_{link}, ct)$ which indicates that a basic triple tr can be associated with a specific context ct. The interpretation of $[[context]] = ([[tr]], [[context_{link}]], [[ct]]) \in U^3 \times U \times U$.

Note that the notion of context spreads to all elements involved in this representation. If we consider only one context, all definitions above can be considered on a specific (unique and implicit) context. In the case that exists more than one context, definitions above can be generalized as follows:

multi-context graph: is an extracted graph with more than one context associated to its triples. For example, given triples tr_i, tr_{i+1}, tr_j and tr_{j+1} in a extracted graph we can have: (i)$(tr_i, context_{link}, ct_1)$, $(tr_{i+1}, context_{link}, ct_1)$, where we have a path associated with context ct_1 (ii) $(tr_j, context_{link}, ct_2)$, $(tr_{j+1}, context_{link}, ct_2)$, and a path associated with context ct_2.

In this case, we can define two new co-reference operators based on the context[6]. In a multi-context graph, given a specific (basic, reified, operational or complex) path, we have: (i) if all basic triples in a path belong to an unique (same) context, the path is **an unique context(basic, reified, operational or complex) path**; (ii) otherwise, we call this path a **multi-context path**.

V. Discussion: SDGs, RDF(S) & Linked Data

The SDG representation was implemented in a SBE graph extraction framework (Graphia)[2][3][4]. 1033 relations (triples) were extracted from 150 sentences from Wikipedia articles. The distribution of a set of sentence features (mapping to the elements of the semantic model) was determined to ensure a heterogeneous sample. The representation was able to cope with all factual sentences in the sample. The set of sentences used are the same used in [3]. While the discussion in [3] focused on the extraction process, this work provides a more in-depth description of the motivations and the construction of a graph extraction model.

The demand to represent complex context structures over facts is the main gap between the current use of RDF(S) and its use to represent structured discourse graphs. Context representation (e.g. temporal, contextual and discourse semantic dependencies) are frequent features in facts extracted from sentences. The centrality of contextual modelling brings reification and named graphs as a fundamental element for the representation of text structures in a vocabulary/ontology independent way, where reifications should become first-class citizens. Additionally dependent triples, i.e. one or more aligned triples in a triple tree, is another example which requires a principled contextual representation. Reification is supported by RDF(S) but discouraged by SPARQL. Under the Linked Data perspective, contextual modelling is also inhibited. Examples of SDG graphs serialized as RDF can be found in [6]. All elements in the semantic model for SDGs can be defined in RDF using a small vocabulary[6] to define the link types.

RDF datasets representing SDGs as proposed in this work should not be put in the same category of traditional (vocabulary-based) datasets. Graphs representing non-normalized discourses respond to a different demand and despite the possible alignment with traditional datasets through entities, they should be separated into a different category. In the SDG context, data can be consumed by applying the navigational queries patterns defined in section IV.B.

[2]http://graphia.dcc.ufrj.br

VI. Conclusion

This work targets filling a gap on the provision of a principled way to represent extracted facts from natural language texts using structured discourse graphs (SDGs). The representation focuses the provision of an entity-centric graph model which is vocabulary/ontology agnostic and which maximizes the capture of complex semantic dependencies present in natural language texts. The representation defines discourse elements over a graph model and provides an algorithmic interpretation approach over the final graph elements. The interplay between the proposed SDGs and RDF(S)/Linked Data was analyzed.

Acknowledgments.: This work has been funded by Science Foundation Ireland (Grant No. SFI/08/CE/I1380 (Lion-2)). João C. P. da Silva is a CNPq Fellow - Science without Borders (Brazil).

References

[1] B. Harrington, P-R. Wojtinnek. Creating a Standardized Markup Language for Semantic Networks. In Proc. of the 5th IEEE Intl. Conference on Semantic Computing, 2011.

[2] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead and O. Etzion. Open Information Extraction from the Web. In Proc. of the Intl. Joint Conference in Artificial Intelligence, 2007.

[3] A. Freitas, D. S. Carvalho, J. C. P. da Silva, S. O'Riain, E. Curry, A Semantic Best-Effort Approach for Extracting Structured Discourse Graphs from Wikipedia. In Proc. of the 1st Workshop on the Web of Linked Entities, (ISWC), 2012.

[4] D. S. Carvalho, A. Freitas, J. C. P. da Silva, Graphia: Extracting Contextual Relation Graphs from Text, In Proc. of the 10th Extended Semantic Web Conference (ESWC), 2013.

[5] V. Presutti et al., Knowledge extraction based on discourse representation theory and linguistic frames. In Proc. of EKAW, 2012.

[6] Complementary material, uri: http://treo.deri.ie/webs.