A Context-Based Model for the Interpretation of Polysemous Terms

Chrisa Tsinaraki, Yannis Velegrakis,
Nadzeya Kiyavitskaya, and John Mylopoulos

Department of Information Engineering and Computer Science (DISI),
University of Trento, Via Sommarive 14, Povo (TN), 38100, Italy
{chrisa,velgias,nadzeya,jm}@disi.unitn.eu

Abstract. The problem of polysemy involves having terms, such as "truck", that refer to multiple concepts in different contexts; and conversely, having the same concept referred to with different names in different contexts. Contexts may be defined along different dimensions, such as language (Italian, English, French, ...), domain (Philosophy, Computer Science, Physics, ...), time (Ancient Greece, 20th century, ...) etc. Given a conceptual model M (aka ontology), a context C and a query Q we motivate and propose algorithms for interpreting all the terms of the query with respect to M and C. We also define and solve the inverse problem: given a set of concepts S which are part of the answer to query Q and a context C, we propose algorithms for choosing terms for all the concepts in S. To illustrate the framework, we use a case study involving a history ontology whose elements are named differently depending on the time period and language of the query.

Keywords: Polysemy, Context, Ontology, Multilingualism.

1 Introduction

The advent of the Internet and the World Wide Web (WWW) have made vast amounts of information – including conceptual models, i.e. ontologies in the Semantic Web jargon – available to all the people of the world. This has brought to the foreground in Information Technology (IT) research an old problem: people of the world speak different languages and, due to this, information needs to be made available to them accordingly.

The general problem of multilingual information sources, such as digital libraries, is an active area of research in Computational Linguistics and Computer Science. We focus in this paper on a smaller problem: polysemy in conceptual models, such as ontologies. A word (or more generally, sign) is polysemous if it means different things in different contexts. Polysemy includes the dual phenomenon of having a concept referred to by different names in different contexts, such as the concept of "ontology" in Philosophy being referred to as "ontology" in English, "ontologia" in Italian and "ontologie" in French and German. Polysemy in the Social Sciences is not just a problem of multilingualism. Terms

R. Meersman et al. (Eds.): OTM 2010, Part II, LNCS 6427, pp. 939–956, 2010.

change meaning over time, topic, and dialect. Again, the term "ontology" means different things in Philosophy and Computer Science (topic) and meant different things in Ancient Greece and the 20^{th} century. These issues are of great concern to historians, for example, who need to take into account all these dimensions that affect polysemy as they analyze a given collection of documents. Unfortunately, polysemy is treated only at the level of locating the synonyms (while working with only one language) and/or the translations (while working with more than one languages) of the terms, taking into account only the spoken language factor [1]; thus, the term context is essentially ignored.

The main objective of this paper is to propose a general mechanism for the interpretation of polysemous terms in environments where a given concept is referred to by different names depending on context. Contexts are defined along several dimensions, consistently with [2]. Given a polysemous conceptual model, a context and a query, we propose algorithms for identifying possible interpretations for the query and also for naming the concepts that are included in the query response.

This research was motivated by, and uses as key case study, a project involving a history ontology used by historians to express queries such as "show me the evolution of the semantics of the term 'biotechnology' in the 20^{th} century". This query can be formalized as: "given a modern name of a concept, identify all the documents related to it in a multilingual archive". The challenge arising from this task is posed by the fact that the terminology changes with time. The users of the history ontology are not aware of the obsolete terms used for the concept in the past in their language; they also have no clue of the possible equivalent terms used for this concept in other languages. For instance, biotechnology was referred at different periods of its development as "biontotechnology", "zymotechnology", "biotechnics", and "biological engineering". Moreover, the history of biotechnology in different countries is very diverse. For example, the translation of this word was equivalent at some point to "biotechnical chemistry" (literal translation from Danish "bioteknisk kemi"). As a result, the lexical translation of the term "biotechnology" is not enough to discover relevant documents written in other languages and in different time periods.

The remainder of the paper is structured as follows: The motivation for our work and the proposed solution are discussed in Section 2, the context-based framework that we have developed for the interpretation of polysemous terms is described in Section 3, a case study that focuses on the formation of biotechnology in the perspective of the History of Science and Technology is provided in Section 4, the related work is presented in Section 5 and the paper concludes in Section 6, where our future research directions are also outlined.

2 Motivation and Solution

Consider the English term "ontology", which has the Italian translation "ontologia". This term was introduced by the ancient Greek philosophers in order to describe a subdomain of metaphysics that was dealing with the philosophical

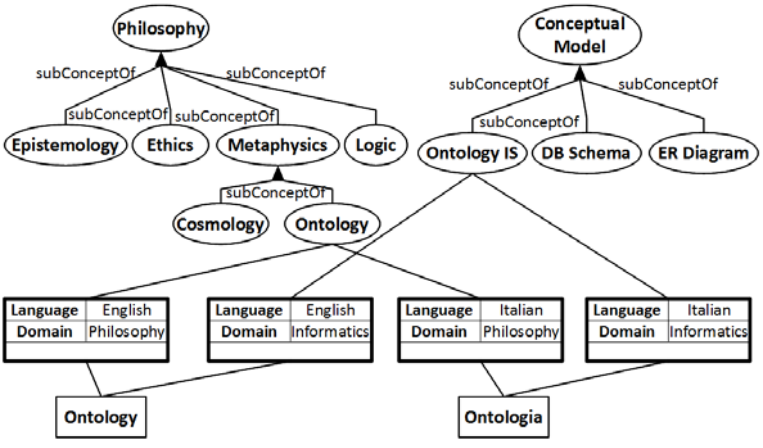

Fig. 1. Context-based associations of the terms "Ontology" and "Ontologia" with the appropriate concepts

study of the nature of being, existence or reality in general, as well as the basic categories of beings and their relations [3]. In addition to its use in Philosophy, the "ontology" term has been recently used in Computer Science in order to describe a formal, explicit specification of a shared conceptualization [4]. Thus, the interpretation of the term "ontology" is highly context-dependent, with the context including both the language and the application domain, as is shown in Fig. 1.

Consider now a digital library that contains material from several different disciplines. This material may be written in any spoken language. In this setting, consider an Italian-speaking informatics researcher that poses a query containing the term "ontologia". The lexical translation of the term in English is not sufficient for retrieving only the relevant material written in English; the user query context, and in particular the application domain, should be specified in order to interpret the query correctly.

In order to satisfy this requirement, we have developed a novel model in which we differentiate between *terms* (i.e. string values used in the data) and *concepts* (i.e. constructs modeling real world concepts and artifacts).

We associate terms to concepts. Each association comes with some confidence and is valid only under certain conditions that are determined through sets of parameters representing *contexts*. For the context representation we use the traditional definition of a vector of N values, the context dimensions [2]. In our application we have chosen 8 dimensions, which we found broadly applicable in real-world scenarios.

The queries are expressed using terms. Based on the query terms, we find the concepts related to them. This is done by taking into account the association context, From the concepts found, we select the terms related to them, again based on the context. Then we retrieve the data that contain these terms.

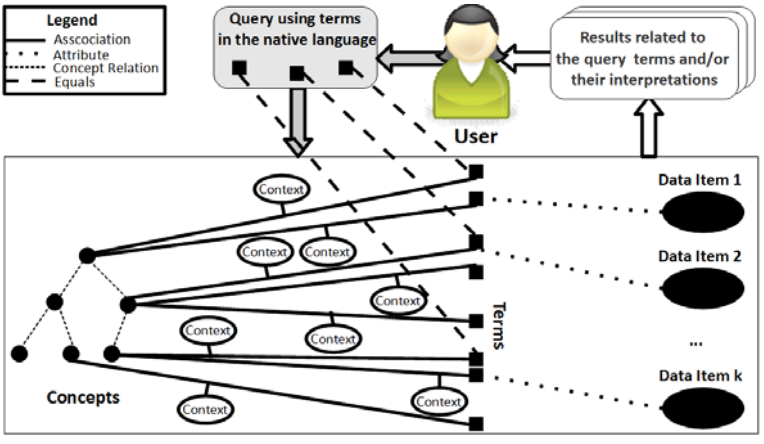

Fig. 2. Generic Cross-lingual Retrieval Use Case

The generic use case supported by our framework is outlined in Fig. 2. As shown, the user poses her queries using terms from her native language. These terms are associated with language-neutral concepts. Since the concepts may be associated with other terms through concept-term associations, the query results may contain any interpretation of the original query terms. Notice that the users may specify (explicitly or implicitly) in their queries some context and receive results related to the query terms under the respective context. This way, the query terms are not interpreted in isolation, but only relative to the query they appear in. Accordingly, if the user in our example specifies the "Informatics" domain in the context of a query containing the "ontologia" term, the documents returned from the digital library are those containing the term "ontologia" as it is used in Informatics or any of its valid interpretations. Our approach is consistent with Gottlob Frege's *Context Principle* [5], a form of semantic holism from Philosophy of Language holding that a philosopher should *"never ... ask for the meaning of a word in isolation, but only in the context of a proposition"*.

3 Context-Based Interpretation Model

We present here the context-based framework that we have developed for the interpretation of polysemous terms. The fundamental concepts of the framework are defined in Section 3.1, the query term interpretation is described in Section 3.2, the model implementation is presented in Section 3.3 and the implementation of the query term interpretation is presented in Section 3.4.

3.1 Fundamental Concepts

Let $Q(t_1, t_2, \cdots, t_n)$ be a query that involves n terms $t_1, t_2, \cdots, t_n$ and $R(t_1, t_2, \cdots, t_n)$ be the query results of $Q(t_1, t_2, \cdots, t_n)$. If there exist valid

interpretations of the query terms, then $\forall\ t_i, i = 1 \cdots n, \exists\ \mathbb{IT}(t_i)$, where $\mathbb{IT}(t_i)$ is the set of the valid interpretations of t_i. If these interpretations should be taken into account, then $R(t_1, t_2, \cdots, t_n)$ should include the results $R'(t'_1, t'_2, \cdots, t'_n)$ of every query $Q'(t'_1, t'_2, \cdots, t'_n)$ with $t'_i \in \mathbb{IT}(t_i)$. If there exist m combinations of the valid interpretations of $t_1, t_2, \cdots, t_n$ and $R'_j(t'_{j1}, t'_{j2}, \cdots, t'_{jn})$ is the query containing the j_{th} combination, then (1) holds.

$$R(t_1, t_2, \cdots, t_n) = R'_1(t'_{11}, t'_{12}, \cdots, t'_{1n}) \cup \cdots \cup R'_m(t'_{m1}, t'_{m2}, \cdots, t'_{mn}) \quad (1)$$

Let s be a concept, t a term and c a context, such that s is associated with t under the context c through an association $a(s, t, c, w)$, where w is a numeric value in the range $[0, 1]$ and describes the strength of the association.

A context c, formally described in (2), is used in order to localize the term-concept associations, while it preserves the compatibility of the reasoning that can be performed in different contexts [6]. According to (2), a context c essentially is a vector of name-value pairs, the *context dimensions*.

$$c\langle d_1 : v_1, \ldots, d_k : v_k\rangle \quad (2)$$

Thus, in the example of Section 2 the term "ontologia" is associated with the "Ontology" concept through an association $a_1(\text{`}Ontology', \text{`}ontologia', c_1\langle language : \text{`}Italian', domain : \text{`}Philosophy'\rangle, 1)$ and with the concept "Ontology IS" through an association $a_2(\text{`}OntologyIS', \text{`}ontologia', c_2\langle language : \text{`}Italian', domain : \text{`}Informatics'\rangle, 1)$.

A *top* value is defined for every context dimension $d_i, (i = 1 \ldots k)$, which includes all the values of the domain of the dimension and is denoted by the $\top$ symbol. If the value v_i of the context dimension d_i of c is not specified, this dimension is assumed to have the $\top$ value. A context $c\langle d_1 : v_1, \ldots, d_k : v_k\rangle$ has the top context value (denoted by $\top$) if it has the $\top$ value in all the dimensions.

A *null* value is also defined for every context dimension $d_i, (i = 1 \ldots k)$, which has the "no value" sense and is denoted by the $\bot$ symbol. A context $c\langle d_1 : v_1, \ldots, d_k : v_k\rangle$ has the null context value (denoted by $\bot$) if it has the $\bot$ value at least in one of its dimensions. For example, the context $c\langle language : \text{`}Italian', domain : \bot\rangle$ has the $\bot$ value.

A partial order is defined for each of the context dimensions, so that their values can be effectively compared. The $\prec$ operator allows detecting if the value v_{ai} of the i^{th} dimension $(i = 1 \ldots k)$ of a context $c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle$ is lower than the value v_{bi} of the i^{th} dimension of a context $c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle$, denoted as $v_{ai} \prec v_{bi}$. For example, $\text{`}1821 - 1936' \prec \text{`}1789 - 1940'$.

Based on the partial order of the context values, a context $c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle$ is more abstract than a context $c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle$, denoted as $c_a \prec c_b$, iff $\forall\ i, (i = 1 \ldots k), v_{ai} \preceq v_{bi}$ and $\exists\ m, (m = 1 \ldots k)$ such that $v_{am} \prec v_{bm}$. As an example, consider two contexts $c_a\langle t : \text{`}1821 - 1936'\rangle$ and $c_b\langle t : \text{`}1789 - 1940'\rangle$, where t is the name of the time dimension of the contexts c_a and c_b. According to the context partial order definition, $c_a \prec c_b$.

The *Greatest Lower Bound* of the values v_{ai} and v_{bi} of the i^{th} dimension of two contexts $c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle$ and $c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle$, denoted as $\mathfrak{glb}(v_{ai}, v_{bi})$, is formally defined in (3).

$$\mathfrak{glb}(v_{ai}, v_{bi}) = g, ((g \preceq v_{ai}) \wedge (g \preceq v_{bi}) \wedge \nexists\, g'((g \prec g') \wedge (g' \preceq v_{ai}) \wedge (g' \preceq v_{bi}))) \quad (3)$$

For example, according to the definition of $\mathfrak{glb}$, we have for the time dimension of the previously defined contexts c_a and c_b that $\mathfrak{glb}(\text{`}1821-1940', \text{`}1789-1936') = \text{`}1821-1936'$.

The *Least Upper Bound* of the values of the i^{th} dimension of two contexts $c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle$ and $c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle$, denoted as $\mathfrak{lub}(v_{ai}, v_{bi})$, is formally defined in (4). For example, for the time dimension of the contexts $c_a\langle t : \text{`}1821-1940'\rangle$ and $c_b\langle t : \text{`}1789-1936'\rangle$ we have $\mathfrak{lub}(\text{`}1821-1940', \text{`}1789-1936') = \text{`}1789-1940'$.

$$\mathfrak{lub}(v_{ai}, v_{bi}) = l, (v_{ai} \preceq l) \wedge (v_{bi} \preceq l) \wedge \nexists\, l'((l' \prec l) \wedge (v_{ai} \preceq l') \wedge (v_{bi} \preceq l')) \quad (4)$$

The *Greatest Lower Bound* $\mathfrak{GLB}(c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle, c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle)$ of two contexts $c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle$ and $c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle$ is a context c', formally defined in (5).

$$\mathfrak{GLB}(c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle, c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle) = c'\langle \mathfrak{glb}(v_{a1}, v_{b1}), \ldots, \mathfrak{glb}(v_{ak}, v_{bk})\rangle \quad (5)$$

For example, assume that t and p are, respectively, the names of the time and place context dimensions. Then, $\mathfrak{GLB}(c_a\langle t : \text{`}1821-1940', p : \text{`}Greece'\rangle, c_b\langle t : \text{`}1789-1936', p : \text{`}Europe'\rangle) = c'\langle t : \text{`}1821-1936', p : \text{`}Greece'\rangle$. Also, $\mathfrak{GLB}(c_a\langle t : \text{`}1821-1940', p : \text{`}Greece'\rangle, c_b\langle t : \text{`}1789-1936', p : \text{`}Canada'\rangle) = c'\langle t : \text{`}1821-1936', p : \perp\rangle = \perp$.

The *Least Upper Bound* $\mathfrak{LUB}(c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle, c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle)$ of two contexts $c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle$ and $c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle$ is a context c', formally defined in (6). For example, $\mathfrak{LUB}(c_a\langle t : \text{`}1821-1940', p : \text{`}Greece'\rangle, c_b\langle t : \text{`}1789-1936', p : \text{`}Europe'\rangle) = c'\langle t : \text{`}1789-1940', p : \text{`}Europe'\rangle$.

$$\mathfrak{LUB}(c_a\langle d_{a1} : v_{a1}, \ldots, d_{ak} : v_{ak}\rangle, c_b\langle d_{b1} : v_{b1}, \ldots, d_{bk} : v_{bk}\rangle) = c'\langle \mathfrak{lub}(v_{a1}, v_{b1}), \ldots, \mathfrak{lub}(v_{ak}, v_{bk})\rangle \quad (6)$$

The framework that we have developed is generic and it does not depend neither on the number of the context dimensions nor on the parameter described by each dimension. From our studies on real-world problems, we suggest that a context c for the interpretation of polysemous query terms comprises of the following $k = 8$ dimensions:

- $d_1 = l$, which represents the language of c;
- $d_2 = p$, which represents the place of c;
- $d_3 = t$, which represents the time period(s) covered by c;
- $d_4 = d$, which represents the application domain of c;
- $d_5 = h$, which represents the historiographical issues (i.e. social conditions, economical issues etc.) that should hold for c to be valid;
- $d_6 = dl$, which represents the dialect of c;

- $d_7 = dt$, which represents the diatype of c (i.e. a language variation, determined by its social purpose [7] like, for example, the specialized language of an academic journal); and
- $d_8 = f$, which represents the formality of c and may take one of the values "Very formal", "Formal", "Neutral", "Informal", "Very informal").

The above listed context dimensions have been identified in our case studies. We selected these dimensions in order to cover: *(a)* The variations of the term semantics in terms of language (d_1), dialect (d_6), social purpose, i.e. diatype, (d_7), formality (d_8) and place (d_2); *(b)* The evolution of the term semantics in time (d_3); and *(c)* The influence of the application domain (d_4) and the historiographical issues (d_5) in the interpretation of polysemous terms.

3.2 Query Term Interpretation

Let $\mathbb{T}$ be the set of all the terms, $\mathbb{S}$ the set of all the concepts, $\mathbb{C}$ the set of all the contexts and $\mathbb{A}$ the set of all the associations. $\mathbb{T}(s) \subseteq \mathbb{T}$ is the set of the terms associated with a concept s and $\mathbb{T}(s, c) \subseteq \mathbb{T}(s)$ is the set of the terms related to s under the c context through an association $a(s, t, c, w)$ with $w \geq w_0$, where $w_0 > 0$ is a threshold value for the strength. For instance, in the example of Section 2, $\mathbb{T}(\text{“}Ontology''\text{)} = \{\text{“}Ontologia'', \text{“}Ontology''\}$ and $\mathbb{T}(\text{“}Ontology', c\langle l : \text{`}Italian'\rangle) = \{\text{“}Ontologia''\}$.

Let also $\mathbb{S}(t) \subseteq \mathbb{S}$ be the set of the concepts associated with a term t. Then, for every pair $\langle t, c\rangle$ exists a (possibly empty) set $\mathbb{S}(t, c)$, comprised of concepts associated with t under c through an association with $w \geq w_0$, as in (7). In addition, for every $s_i \in \mathbb{S}(t, c)$ exists a (possibly empty) set $\mathbb{T}_i(s_i, c_i)$, comprised of terms associated with s_i through an association $a_i(s_i, t, c_i, w_i)$ with $w_i \geq w_0$, as in (8). Notice that $|\mathbb{P}|$ is the cardinality of a set $\mathbb{P}$. In our example, for instance, we have $\mathbb{S}(\text{“}Ontologia''\text{)} = \{\text{“}Ontology'', \text{“}OntologyIS''\}$ and $\mathbb{S}(\text{“}Ontologia'', c\langle l : \text{`}Italian', d : \text{`}Informatics'\rangle) = \{\text{“}OntologyIS''\}$.

$$\forall\ \langle t, c\rangle, t \in \mathbb{T}, c \in \mathbb{C}, \exists\ \mathbb{S}(t, c') \subseteq \mathbb{S}, c' \in \mathbb{C}, c' \preceq c \tag{7}$$

$$\forall\ s_i \in \mathbb{S}(t, c), \exists\ \mathbb{T}_i(s_i, c_i) \subseteq \mathbb{T}, c_i \in \mathbb{C}, c_i \preceq c, i = 1 \ldots M, M = |\mathbb{S}| \tag{8}$$

Let $Q(t_1, t_2, \ldots, t_h, c)$ be a generic query that involves, under the c context, h terms $t_1, \ldots, t_h$ and $R(t_1, \ldots, t_h, c)$ be the results of $Q(t_1, \ldots, t_h, c)$. Let also $Q'(s_1, \ldots, s_p, c')$ be a generic query that involves, under the c' context, p concepts $s_1, \ldots, s_p$ and $R'(s_1, \ldots, s_p, c')$ be the results of $Q'(s_1, \ldots, s_p, c')$. Then (9) and (10) hold and $R(t_1, \ldots, t_h, c)$ is calculated as in (11).

$$\forall\ \langle t_i, c\rangle, t_i \in \mathbb{T}, c \in \mathbb{C}, \exists\ \mathbb{S}_i(t_i, c_i) \subseteq \mathbb{S}, c_i \in \mathbb{C}, c_i \preceq c, i = 1 \ldots h \tag{9}$$

$$\begin{aligned} &\forall\ s_{ij} \in \mathbb{S}_i(t_i, c_i), \exists\ \mathbb{T}_{ij}(s_{ij}, c_{ij}) \subseteq \mathbb{T}, c_{ij} \in \mathbb{C}_i, c_{ij} \preceq c_i, \\ &i = 1 \ldots h, j = 1 \ldots M_i, M_i = |\mathbb{S}_i| \end{aligned} \tag{10}$$

$$R(t_1,\ldots,t_h,c) = R'_1(s_{11},\ldots,s_{1h},c'_1)\cup\ldots\cup R'_T(s_{T1},\ldots,s_{Th},c'_h),\\ s_{ij}\in\mathbb{S}_i(t_i,c_i), c'_i=\mathfrak{GLUB}(c_{i1},\ldots,c_{ih}), c'_i\neq\perp\\ i=1\ldots h, T\leq h\cdot max(M_1,M_2,\ldots M_h) \tag{11}$$

Finally, $R'_i(s_{i1},\ldots,s_{ih},c'_i)$ is calculated as in (12).

$$R'_i(s_{i1},\ldots,s_{ih},c'_i) = R_{i1}(t_{11},\ldots,t_{1h},c_{i1})\cup\ldots\cup R_{ih}(t_{h1},\ldots,t_{hh},c_{ih})\\ t_{ij}\in\mathbb{T}_i(s_{ij},c'_i), c'_{ij}=\mathfrak{GLUB}(c_{11},\ldots,c_{hh}), c'_{ij}\neq\perp\\ i=1\ldots h, j=1\ldots h \tag{12}$$

For instance, using the above notation the query of our example is expressed as $Q(\text{“}Ontologia'', c\langle l:\text{‘}Italian', d:\text{‘}Informatics'\rangle)$ and the query results are $R(\text{“}Ontologia'', c\langle l:\text{‘}Italian', d:\text{‘}Informatics'\rangle)$. From (11) we have (13) and from (12) we have (14). Thus, (15) holds.

$$R(\text{“}Ontologia'',\ c\langle l:\text{‘}Italian', d:\text{‘}Informatics'\rangle) =\\ R'(\text{“}Ontology'', c\langle l:\text{‘}Italian', d:\text{‘}Informatics'\rangle) \tag{13}$$

$$R'(\text{“}Ontology'',\ c\langle l:\text{‘}Italian', d:\text{‘}Informatics'\rangle) =\\ R(\text{“}Ontologia'', c\langle l:\text{‘}Italian', d:\text{‘}Informatics'\rangle)\cup\\ R(\text{“}Ontology'', c\langle l:\text{‘}English', d:\text{‘}Informatics'\rangle) \tag{14}$$

$$R(\text{“}Ontologia'',\ c\langle l:\text{‘}Italian', d:\text{‘}Informatics'\rangle) =\\ R(\text{“}Ontologia'', c\langle l:\text{‘}Italian', d:\text{‘}Informatics'\rangle)\cup\\ R(\text{“}Ontology'', c\langle l:\text{‘}English', d:\text{‘}Informatics'\rangle) \tag{15}$$

3.3 Model Implementation

The proposed model has been realized in a dataspace in the context of the TRENDS system [8]. The dataspace data model is flexible enough to consider data of different types, such as semi-structured, relational or RDF. In contrast to data models like the relational that are based on some structural notion, e.g., the notion of a tuple, the dataspace model is centered around the notion of *entity*, which is used to model any real-world object (web document, relational tuple, image, spreadsheet record). As it is typically the case in real world scenarios, despite the fact that entities can be clustered into groups with similar characteristics, it is rarely the case that all the entities in the same group will fully conform to the same strict specification. As such, the basic model imposes no restrictions on the structure and characteristics of the entities [9]. Only a unique identifier is stored for every object regardless of the object type.

We assume the existence of an infinite set of entities $\mathcal{E}$, of names $\mathcal{N}$ and of atomic values $\mathcal{V}$ containing values such as integers, strings, reals, etc.

Definition 1. *A* dataspace *is a tuple* $\langle E,G\rangle$ *where* E *and* G *are finite sets with* $E\subseteq\mathcal{E}$ *and* $G\subseteq E\times\mathcal{N}\times\{E\cup\mathcal{V}\}$. *An* attribute *is a pair* $\langle n,v\rangle$ *with the* $n\in\mathcal{N}$ *being referred to as the* attribute name *and the* $v\in\mathcal{E}\cup\mathcal{V}$ *as the* attribute value. *The attributes of an entity* e *in a dataspace* $\langle E,G\rangle$ *is the set* $A(e)=\{\langle n,v\rangle \mid \langle e,n,v\rangle\in G\}$.

The data model contains constructs to support schema information when this is required. In particular, certain entities in the dataspace can be used to describe the structure of a set of entities. These entities are called *classes*. To declare that the structure of an element e is described by a class c, an attribute *type* with value c is added in e. To identify the entities that can serve as classes, we assume by default in every dataspace the existence of a special entity *Class* and we require that every class entity has an attribute *type* with value *Class*. Furthermore, hierarchies among classes can be defined using a predefined attribute *subtype*.

The implementation of the context-based model for the interpretation of polysemous query terms is built upon three types of special entities: The *concepts*, the *terms* and the *concept-term associations*.

Definition 2. *A concept is a dataspace entity* s*, which belongs to the* Concept *class and has the* $A(s)$ *set of attributes.*

$$A(s) = \{\langle type : `Concept'\rangle, \langle id : sid\rangle, \langle name : sn\rangle, \langle concept\_relation : r\rangle\} \tag{16}$$

The *type* attribute associates s with the *Concept* class to which s belongs, the *id* attribute identifies s, the *name* attribute represents the name of s and the *concept_relation* attribute represents a relationship between concepts (like, for example, sub_concept_of, refines, generalizes ...) in which s participates.

The value of the *concept_relation* attribute is a dataspace entity r of *ConceptRelation* type, which has the $A(r)$ set of attributes formally described in (17). The *type* attribute associates r with the *ConceptRelation* class to which r belongs, the *id* attribute identifies r, the *source* attribute represents the source of the relationship, the *target* attribute represents the target of the relationship and the *concept_relation_type* attribute specifies the relationship type.

$$\begin{aligned} A(r) = \{&\langle type : `ConceptRelation'\rangle, \langle id : rid\rangle, \langle source : src\rangle, \langle target : trg\rangle, \\ &\langle concept\_relation\_type : t\rangle\} \end{aligned} \tag{17}$$

For instance, the "Ontology IS" concept in the example of Section 2 is represented by a dataspace entity s_1 that has the $A(s_1)$ set of attributes.

$$\begin{aligned} A(s_1) = \{&\langle type : `Concept'\rangle, \langle id : `OntologyIS'\rangle, \langle name : `OntologyIS'\rangle, \\ &\langle concept\_relation : `r_1'\rangle\} \end{aligned} \tag{18}$$

Notice that the "Ontology IS" concept in our example is associated with the "Conceptual Model" concept through a relation of type "subConceptOf". This relation is represented by the dataspace entity r_1 that has the $A(r_1)$ set of attributes.

$$\begin{aligned} A(r_1) = \{&\langle type : `ConceptRelation'\rangle, \langle concept\_relation\_type : `subConceptOf'\rangle, \\ &\langle id : `r_1'\rangle, \langle source : `OntologyIS'\rangle, \langle target : `ConceptualModel'\rangle\} \end{aligned} \tag{19}$$

Definition 3. *A term is a dataspace entity* t*, which belongs to the* Term *class and has the* $A(t)$ *set of attributes.*

$$A(t) = \{\langle type : `Term'\rangle, \langle id : tid\rangle, \langle name : tn\rangle\} \tag{20}$$

The *type* attribute associates t with the *Term* class to which t belongs, the *id* attribute identifies t and the *name* attribute represents the name of t.

For instance, the "Ontologia" term in our example is represented by a dataspace entity t_1 that has the $A(t_1)$ set of attributes.

$$A(t_1) = \{\langle type : \text{`}Term'\rangle, \langle id : \text{``}Ontologia''\rangle, \langle name : \text{``}Ontologia''\rangle\} \quad (21)$$

Definition 4. *A concept-term association is a dataspace entity* cta*, which belongs to the* CTAssociation *class and has the* $A(cta)$ *set of attributes.*

$$A(cta) = \{\langle type : \text{`}CTAssociation'\rangle, \langle id : ctaid\rangle, \langle concept : s\rangle, \langle term : t\rangle, \langle confidence : w\rangle, \langle definition : d\rangle, \langle context : c\rangle\} \quad (22)$$

The *type* attribute associates *cta* with the *CTAssociation* class to which *cta* belongs, the *id* attribute identifies *cta*, the *concept* attribute links *cta* to the participating concept s and the *term* attribute links *cta* to the participating term t. The confidence of *cta* is represented by the *confidence* attribute, the textual definition of t under *cta* is represented by the *definition* attribute and the *context* attribute specifies the association context of *cta*.

For instance, the "Ontology IS" concept in our example is associated with the "Ontologia" term through a concept-term association. This association is represented by a dataspace entity cta_1 that has the $A(cta_1)$ set of attributes.

$$A(cta_1) = \{\langle type : \text{`}CTAssociation'\rangle, \langle id : \text{`}cta_1'\rangle, \langle concept : \text{`}s_1'\rangle, \langle term : \text{`}t_1'\rangle, \langle confidence : \text{`}1'\rangle, \langle definition : \text{`}...'\rangle, \langle context : \text{`}c_1'\rangle\} \quad (23)$$

The value of the *context* attribute is a dataspace entity c of *Context* type, which has the $A(c)$ set of attributes, formally described in (24).

$$A(c) = \{\langle type : \text{`}Context'\rangle, \langle id : cid\rangle, \langle language : l\rangle, \langle place : p\rangle, \langle time : t\rangle, \langle domain : d\rangle, \langle historiographical\_issues : h\rangle, \langle dialect : dl\rangle, \langle diatype : dt\rangle, \langle formality : f\rangle\} \quad (24)$$

The *type* attribute associates c with the *Context* class to which c belongs, the *id* attribute identifies c and the rest of the (optional) attributes essentially are the context dimensions; In particular, the *language* attribute specifies the language of c, the *place* attribute specifies the place of c, the *time* attribute specifies the time of c, the *domain* attribute specifies the domain of c, the *historiographical_issues* attribute specifies the historiographical issues of c, the *dialect* attribute specifies the dialect of c, the *diatype* attribute specifies the diatype of c and the *formality* attribute specifies the formality of c.

For example, the context of the cta_1 concept-term association is represented by a dataspace entity c_1 that has the $A(c_1)$ set of attributes.

$$A(c_1) = \{\langle type : \text{`}Context'\rangle, \langle id : \text{`}c_1'\rangle, \langle language : \text{`}Italian'\rangle, \langle domain : \text{`}Informatics'\rangle\} \quad (25)$$

3.4 Implementation of Query Term Interpretation

In our working environment [10][11], a query is described by a rule. A rule consists of two parts: the *head* and the *body*. Each part is a conjunction of atoms

(or subgoals). There are two kinds of atoms: the arithmetic and the entity atoms. An arithmetic atom is a Boolean condition that involves variables and constant values, i.e., $x \leq 10$ or $x=y$. Arithmetic atoms can appear in the body of the rule but not in the head. An entity atom is of the form $e(n_1{:}v_1,\ n_2{:}v_2,\ \ldots,\ n_k{:}v_k)$, where e, n_i and v_i are variables or constants. Variables appearing outside the parenthesis in an atom, like the variable e, are called entity variables and can be bound only to dataspace entities. Variables like the n_1, n_2, ... can be bound only to attribute names, and variables like the v_1, v_2, ... can be bound either to atomic values or to entities. Given a binding of the variables e, n_i, v_i to e^b, n_i^b and v_i^b, respectively, for every $i{=}1..k$, the entity atom $e(n_1{:}v_1,\ n_2{:}v_2,\ \ldots,\ n_k{:}v_k)$ is said to be true if there is an entity e^b in the dataspace that has attributes $\langle n_i^b{:}v_i^b\rangle$, for every $i{=}1..k$.

The atoms in the head of a query are true, if the all the atoms in the body of the query are true. When the body of the query is evaluated to true, a set of entities and attributes as described by the head of the query are returned. Thus, the result of a query in our model is itself a dataspace which makes the query language closed over the set of all possible dataspaces and allowing the composition of queries. For readability purposes and to reduce the number of arithmetic operations, variables may be shared across atoms or be replaced by constants. Furthermore, atoms may be used in attributes as values to simulate nesting.

In order to perform the context-based query term interpretation in our dataspace working environment, the original queries should be rewritten in such a way that the required attribute values and/or names can also be matched by their interpretations in a given context.

Let Q be a generic dataspace query, of the form shown in expression (26), where a_i $(i = 1 \ldots n)$ are attribute names and v_i are attribute values. The query context is $c(language : l, place : p, time : t, domain : d, dialect : dl, diatype : dt, formality : f, historiographical\_issues : h)$. The semantics of the query Q is that the results should include all the entities having the value v_i in the attribute a_i. If v_i is a literal value, the results should also include all the entities having an interpretation v_i as value of the a_i attribute. In addition, if a_i is a literal value, the results should also include all the entities with the desired value in an attribute having as name an interpretation of a_i. For every query result x, the context c' under which x is valid is also specified.

$$\begin{aligned} \$\mathbf{x}(context : \$c') :- \ & \$\mathbf{x}(a_1 : v_1, a_2 : v_2, \cdots, a_n : v_n), \$\mathbf{c}(language : l, \\ & place : p, time : t, domain : d, dialect : dl, \\ & diatype : dt, formality : f, historiographical\_issues : h) \end{aligned} \tag{26}$$

In our implementation, Q is rewritten in a way that exploits the concept-term associations in order to allow the required attribute values and names to be matched by their interpretations. The equivalent query Q' is shown in expression (27). A weight w is also specified for every result of Q', which is calculated on the basis of the strengths of the concept-term associations.

$$
\begin{aligned}
&\$\mathbf{x}(\$w : \$w_{sv1} \cdot \$w_{sk1} \cdot \$w_{av1} \cdot \$w_{ak1} \cdot \ldots \cdot \$w_{svn} \cdot \$w_{skn} \cdot \$w_{avn} \cdot \$w_{akn},\\
&\quad context : \$c') :- \$\mathbf{x}(\$key_{a1} : \$key_{v1}, \ldots, \$key_{an} : \$key_{vn}),\\
&\quad \$\mathbf{y_{v1}}(concept : \$s_{v1}, type : `CTAssociation', term : v_1, confidence : \$w_{sv1},\\
&\quad context : \$c_{sv1}), \ldots, \$\mathbf{y_{vn}}(concept : \$s_{vn}, type : `CTAssociation',\\
&\quad term : v_n, confidence : \$w_{svn}, context : \$c_{svn}),\\
&\quad \$\mathbf{z_{v1}}(concept : \$s_{v1}, type : `CTAssociation', term : \$key_{v1},\\
&\quad confidence : \$w_{sk1}, context : \$c_{sk1}), \ldots, \$\mathbf{z_{vn}}(concept : \$s_{vn},\\
&\quad type : `CTAssociation', term : \$key_{vn}, confidence : \$w_{skn}, context : \$c_{skn}),\\
&\quad \$\mathbf{y_{a1}}(concept : \$s_{a1}, type : `CTAssociation', term : a_1, confidence : \$w_{av1},\\
&\quad context : \$c_{av1}), \ldots, \$\mathbf{y_{an}}(concept : \$s_{a1}, type : `CTAssociation',\\
&\quad term : a_n, confidence : \$w_{avn}, context : \$c_{avn})\\
&\quad \$\mathbf{z_{a1}}(type : `CTAssociation', term : \$key_{a1}, concept : \$s_{a1},\\
&\quad confidence : \$w_{ak1}, context : \$c_{ak1}), \ldots, \$\mathbf{z_{an}}(type : `CTAssociation',\\
&\quad term : \$key_{an}, concept : \$s_{an}, confidence : \$w_{akn}, context : \$c_{akn}),\\
&\quad \$\mathbf{c}(language : l, place : p, time : t, domain : d, dialect : dl, diatype : dt,\\
&\quad formality : f, historiographical\_issues : h)\\
&\quad \mathbf{WITH}\ (\$c_{sv1} \preceq c) \wedge \ldots \wedge (\$c_{svn} \preceq c) \wedge (\$c_{av1} \preceq c) \wedge \ldots \wedge (\$c_{avn} \preceq c)\\
&\quad \wedge (\$c' \neq \bot) \wedge (\$c' = \mathfrak{GLB}(\$c_{sv1}, \ldots, \$c_{svn}, \$c_{av1}, \ldots, \$c_{avn},\\
&\quad \$c_{sk1}, \ldots, \$c_{skn}, \$c_{ak1}, \ldots, \$c_{akn}))
\end{aligned}
\tag{27}
$$

Consider, as an example, the query of our motivating example, which should retrieve all the objects that contain the keyword "ontologia", assuming that: (a) The entity keywords are stored in the *keyword* attribute; (b) The user specifies the Italian term "ontologia"; and (c) The user is interested only in the entities that are referring to the term "ontologia" as it is used in the informatics domain. The query has the form Q_a shown in expression (28), which will retrieve all the entities having "ontologia" as value of the *keyword* attribute, in a context $c(language : `Italian', domain : `informatics')$. In our implementation, Q_a is rewritten in the form of Q'_a shown in expression (29).

$$
\begin{aligned}
&\$\mathbf{x}(context : \$c') :- \$\mathbf{x}(keyword : `ontologia'),\\
&\qquad\qquad \$\mathbf{c}(language : `Italian', domain : `informatics')
\end{aligned}
\tag{28}
$$

$$
\begin{aligned}
&\$\mathbf{x}(\$w : \$w_v \cdot \$w_k, context : \$c') :- \$\mathbf{x}(keyword : \$key),\\
&\quad \$\mathbf{y}(type : `CTAssociation', concept : \$s, term : `ontologia',\\
&\quad confidence : \$w_v, context : \$c_v), \$\mathbf{z}(type : `CTAssociation', concept : \$s,\\
&\quad term : \$key, confidence : \$w_k, context : \$c_k), \$\mathbf{c}(language : `Italian',\\
&\quad domain : `informatics')\\
&\quad \mathbf{WITH}\ (\$c_v \preceq c) \wedge (\$c' = \mathfrak{GLB}(\$c_v, \$c_k)) \wedge (\$c' \neq \bot)
\end{aligned}
\tag{29}
$$

Consider now a query Q_b, which assumes that: (a) The entity keywords are stored in an attribute that has as name an interpretation of "keyword"; (b) The user specifies the English term "ontology"; and (c) The user is interested only in the

entities that are referring to the term "ontology" as it is used in the informatics domain. Q_b has the form of (30) and it is rewritten as in expression (31).

$$\begin{aligned} \$\mathbf{x}(context : \$c') :- \ & \$\mathbf{x}(`keyword' : `ontology'), \\ & \$\mathbf{c}(language : `English', domain : `informatics') \end{aligned} \tag{30}$$

$$\begin{aligned} & \$\mathbf{x}(\$w : \$w_{sv} \cdot \$w_{sk} \cdot \$w_{av} \cdot \$w_{ak}, context : \$c') :- \$\mathbf{x}(\$key_a : \$key_v), \\ & \quad \$\mathbf{y_v}(type : `CTAssociation', concept : \$s_v, term : `ontology', \\ & \quad confidence : \$w_{sv}, context : \$c_{sv}), \$\mathbf{z_v}(type : `CTAssociation', \\ & \quad concept : \$s_v, term : \$key_v, confidence : \$w_{sk}, context : \$c_{sk}), \\ & \quad \$\mathbf{y_a}(type : `CTAssociation', concept : \$s_a, term : `keyword', \\ & \quad confidence : \$w_{av}, context : \$c_{av}), \\ & \quad \$\mathbf{z_a}(type : `CTAssociation', concept : \$s_a, term : \$key_a, confidence : \$w_{ak}, \\ & \quad context : \$c_{ak}), \$\mathbf{c}(language : `English', domain : `informatics') \\ & \quad \textbf{WITH } (\$c_{sv} \preceq c) \wedge (\$c_{av} \preceq c) \wedge (\$c' \neq \bot) \wedge \\ & \quad (\$c' = \mathfrak{GLB}(\$c_{sv}, \$c_{av}, \$c_{sk}, \$c_{ak})) \end{aligned} \tag{31}$$

4 Use Case

To illustrate the application of our approach, we present a case study which focuses on the formation of biotechnology from the perspective of History of Science and Technology, introduced earlier in section 1. In order to answer a user query on biotechnology, we need to identify all the documents related to the term of interest in the underlying archive. This is not a trivial task, given that during the 20^{th} century the notion of biotechnology has undergone many terminological and conceptual changes, as discussed in the seminal essay of Robert Bud [12]. Those changes must be modeled appropriately in order to be able to identify the documents related to the term "biotechnology" in a given query context that may include the time period, country etc.

In particular, the word "biotechnology" was first introduced in 1917 by a Hungarian agricultural engineer, Karl Ereky, to cover the area of technology associated with the living beings.

However, the origins of biotechnology as a field of study go back to the late 19^{th} century in relation to the field of zymotechnology that concerns industrial fermentation and brewing techniques. In different countries different terms were used with a similar meaning, e.g., "biontotechnik", "biotechnik", "biotechnics" and others. In the 1960s, biotechnology came to connote the environment friendly technological orientation rather than a specific technology. Thus, the definition of biotechnology had been very fluid till the 1970s, when it finally obtained its current meaning based on its links to microbiology and genetics. In this setting, conventional query answering mechanisms would return the documents explicitly mentioning the "biotechnology" keyword, while leaving out the rest of the relevant resources which might use, for instance, the obsolete term "zymotechnology" instead.

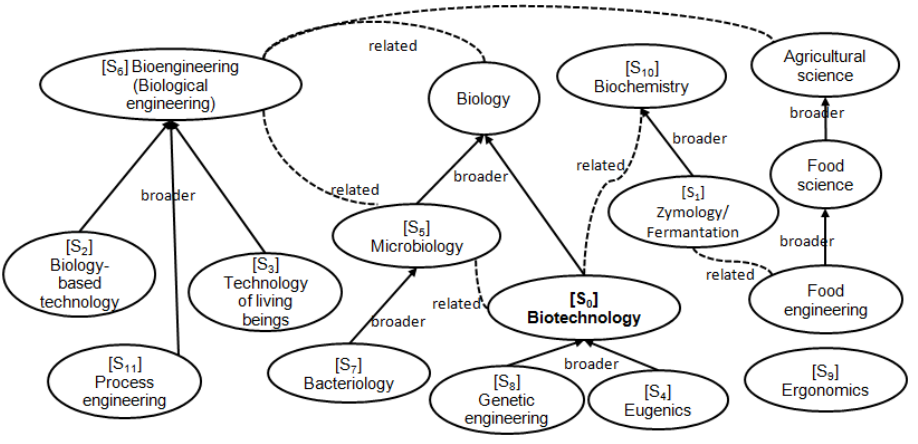

Fig. 3. An approximate model of biotechnology-related fields

See a fragment of the ontology representing the contemporary understanding of the relevant scientific fields in Fig. 3 and the record of the main historical milestones in Table 1.

If we consider our initial query "show me the evolution of the semantics of the term 'biotechnology' in the 20^{th} century" (see Section 1), the proposed context-based model for the interpretation of polysemous terms comes in handy to comprehensively answer such historical queries. In particular, the query is initially expressed, using the term 'biotechnology' in English, as in (32) and is then rewritten as shown in (33). The later will return any valid interpretation of the term 'biotechnology' that was used during the 20^{th} century in any language together with the context in which this interpretation is valid.

$$\begin{aligned} \$\mathbf{x}(context : \$c') :- \; & \$\mathbf{x}(keyword : `biotechnology'), \\ & \$\mathbf{c}(language : `English', time = `20^{th} century') \end{aligned} \tag{32}$$

$$\begin{aligned} \$\mathbf{x}(\$w : \$w_v \cdot \$w_k, context : \$c') :- \; & \$\mathbf{x}(keyword : \$key), \\ & \$\mathbf{y}(type : `CTAssociation', concept : \$s, term : `biotechnology', \\ & confidence : \$w_v, context : \$c_v), \\ & \$\mathbf{z}(type : `CTAssociation', concept : \$s, term : \$key, confidence : \$w_k, \\ & context : \$c_k), \$\mathbf{c}(language : `English', time = `20^{th} century') \\ & \mathbf{WITH}\ (\$c_v \preceq c) \wedge (\$c' = \mathfrak{GLB}(\$c_v, \$c_k)) \wedge (\$c' \neq \bot) \end{aligned} \tag{33}$$

Consider now that the initial query is modified in order to show the recent (after 1970) evolution of the term "biotechnology". The new query is initially expressed as in (34) and is then rewritten as shown in (35). The later will return only the valid interpretations of the term 'biotechnology' used during from 1970 together with the context in which this interpretation is valid.

$$\begin{aligned} \$\mathbf{x}(context : \$c') :- \; & \$\mathbf{x}(keyword : `biotechnology'), \\ & \$\mathbf{c}(language : `English', time = `1970 - now') \end{aligned} \tag{34}$$

Table 1. Terms and related concepts in the history of biotechnology

Term	Concepts	Concept IDs	Context		
			Language	Place	Time (uncertain dates in parentheses)
Bioteknisk kemi	Fermentation physiology (industrial fermentation, agriculture)	S_1	Danish	Denmark	1915–1945
Biotechnologie	Biology-based technology	S_2	Hungarian	Hungary	1917–1945
Biotechnologie	Biology-based technology	S_2	German	Germany	1920–1945
Biontotechnik	Technology of living beings	S_3	German	Germany	1901–1945
Biotechnik	Technology for human improvement (eugenics, social biology)	S_4	German	Germany	1911–1945
Zymotechnology	Brewing technique (industrial fermentation)	S_1	English	UK	1900–1918
Biotechnology	Applied microbiology, brewing	S_1, S_5	English	UK	1918–1939
Biotechnology	Applied microbiology, brewing	S_1, S_5	English	USA	1918–1939
Biotechnology	Biological technology for human improvement (eugenics, social biology)	S_4	English	UK	1936–(1945)
Biotechnics	Technology based on biology	S_2	English	UK	1915–(1962)
Biotechnics	(Bio)engineering	S_6	English	USA	1934–1945
Biological Engineering	(Bio)engineering	S_6	English	USA	1936–(1945)
Biotechnology	Ergonomics (Biologically compatible technology)	S_9	English	USA	1946–1972
Biotechnology	Modern definition	S_0={S_1, S_4, S_5, S_6, S_8}	English	USA	1975–Now
Biotechnik	Biology-based technologies, brewing	S_1, S_2	Swedish	Sweden	1943 – late 1950s
Biotechnik	Biology-based technologies, brewing, bacteriology, eugenics, microbiology	S_1, S_2, S_4, S_5, S_7	Swedish	Sweden	late 1950s – Now
Biotechnologi	Ergonomics	S_9	Swedish	Sweden	1950–(1975)
Biotechnologi	Modern definition	S_0={S_1, S_4, S_5, S_6, S_8}	Swedish	Sweden	(1975)–Now
Biotechnologie	Microbiology, biochemistry, biotechnical chemistry/fermentation, process engineering	S_1, S_5, S_{10}, S_{11}	German	Germany	1974–1980
Biotechnologie	Modern definition	S_0={S_1, S_4, S_5, S_6, S_8}	German	Germany	1980–Now

$$
\begin{aligned}
&\$\mathbf{x}(\$w : \$w_v \cdot \$w_k, context : \$c') :- \$\mathbf{x}(keyword : \$key),\\
&\quad \$\mathbf{y}(type : `CTAssociation', concept : \$s, term : `biotechnology',\\
&\quad confidence : \$w_v, context : \$c_v),\\
&\quad \$\mathbf{z}(type : `CTAssociation', concept : \$s, term : \$key, confidence : \$w_k,\\
&\quad context : \$c_k), \$\mathbf{c}(language : `English', time = `1970 - now')\\
&\quad \textbf{WITH } (\$c_v \preceq c) \wedge (\$c' = \mathfrak{GLB}(\$c_v, \$c_k)) \wedge (\$c' \neq \bot)
\end{aligned}
\tag{35}
$$

5 Related Work

The context-based query term interpretation framework that we have presented in this paper relates with two main research disciplines: *(a)* Systems that support multilingualism; and *(b)* Context-based systems.

Multilingualism. The research works that support multilingualism are usually based on the association of the conceptual model elements (i.e. entities, classes, ...) with terms describing them in different languages. This may be done either by keeping the multilingual labels inside the conceptual model elements [13] or by mapping them to sets of synonym terms [1][14][15]. This approach does not usually allow the association of the conceptual model elements to more than one sets of synonyms. There exist, though, some variations of this approach, based on the alignment of conceptual models built using terms from different spoken languages [16][17][18]. If the alignment involves more than two languages, a pivot language is utilized (usually English), in which all the concepts of the application domain are represented [19].

An important disadvantage of the aforementioned approaches is the limited (or even altogether lack of) context utilization for the disambiguation of the terms used in different languages; only the dubbed context is used in some research works for the disambiguation of the text-translation process [20][21]. In this sense, our context-based framework proposes a new research direction in the multilingualism domain that allows for the more reliable interpretation of the query terms in cross-lingual retrieval. The only approaches that have some similarities with our work are: (a) [22], where the context is used in order to associate terms with concepts, is the approach of the ontological engineering framework DOGMA. In that work, though, the context is not structured and is intended to be used by human readers; and (b) [23], where a domain is associated with every term in the Wordnet lexical database. This is a special case in our context-based model, where the domain is just one of the context dimensions.

Context. The notion of context plays a central role in the framework that we have developed for the interpretation of polysemous query terms. The context model used in our framework is compatible with well-accepted general purpose context models [2], [24], while it is applied in a domain that has not yet benefited from the utilization of context. In particular, both our approach and the aforementioned ones model the context as a set of values, the context dimensions. In addition, we adopt the idea of the partial order of the values of the context dimensions from [2], while we also specify a null context value and the $\mathfrak{LUB}$ and $\mathfrak{GLB}$ context operations, that are compatible, respectively, with the empty context value and the union and intersection context operations that were proposed in [24]. Moreover, our context management approach has similarities with [25], where the context is also handled as a first-class entity (concept or instance). Last, but not least, our use of weights for associations is adopted from Computational Linguistics [26], where has been research on the selection among synonyms in different contexts.

6 Conclusions and Future Work

We have proposed a general-purpose mechanism for dealing with the interpretation of polysemous terms. Our proposal adopts ideas from a number of sources, including multilingual ontologies, contexts and Computational Linguistics. The mechanism has been implemented and is used in a special-purpose digital library founded on a history ontology for renewable energy and biotechnology in the context of the EU project PAPYRUS.

Our future research includes the extensive evaluation of our framework in cross-lingual retrieval. We also intend to explore other potential applications of the proposed framework, both in digital libraries and apart from the digital library domain, that need to correctly interpret polysemous terms in a multi-contextual setting.

Acknowledgement

We acknowledge funding for this research from the PAPYRUS project (ICT-215874). The authors would like to thank our colleagues working on the project, especially our historian colleagues, for introducing us to the complexities of historical research.

References

1. Dong, X., Halevy, A.: Indexing dataspaces. In: Proc. of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 43–54 (June 11- 14, 2007)
2. Bolchini, C., Schreiber, F.A., Tanca, L.: A methodology for a very small data base design. Information Systems 32(1), 61–82 (2007)
3. The Wikipedia Free Encyclopedia, `http://www.wikipedia.org`
4. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5/2, 199–220 (1993)
5. Frege, G.: The Foundations of Arithmetic (EN Transl. by J.L.Austin), 2nd revised edn. (1884/1980)
6. Ghidini, C., Giunchiglia, F.: Local models semantics, or contextual reasoning=locality+compatibility. Artificial Intelligence 127(2), 221–259 (2001)
7. Gregory, M.: Aspects of varieties differentiation. Journal of Linguistics 3, 177–197 (1967)
8. Bykau, S., Kiyavitskaya, N., Tsinaraki, C., Velegrakis, Y.: Bridging the gap between heterogeneous and semantically diverse content of different disciplines. In: Proc. of the 2010 DEXA Workshop (FlexDBIST) (August 30 - September 4, 2010)
9. Dong, X., Halevy, A., Madhavan, J.: Reference reconciliation in complex information spaces. In: Proc. of the 2005 ACM SIGMOD Int. Conf. on Management of Data, pp. 85–96. ACM, New York (2005)
10. Rizzolo, F., Velegrakis, Y., Mylopoulos, J., Bykau, S.: Modeling concept evolution: A historical perspective. In: Laender, A.H.F. (ed.) ER 2009. LNCS, vol. 5829, pp. 331–345. Springer, Heidelberg (2009)
11. Presa, A., Velegrakis, Y., Rizzolo, F., Bykau, S.: Modeling associations through intensional attributes. In: Laender, A.H.F. (ed.) ER 2009. LNCS, vol. 5829, pp. 315–330. Springer, Heidelberg (2009)

12. Bud, R.: Biotechnology in the twentieth century. Social Studies of Science 21, 415–457 (1991)
13. Segev, A., Gal, A.: Egovernment policy evaluation support using multilingual ontologies. In: Proc. of 1st Int. Conf. on Interoperability of eGovernment Services (eGovInterop 2005) (February 23-24, 2005)
14. Kerremans, K., Temmerman, R.: Towards multilingual, termontological support in ontology engineering. In: Proc. of Termino 2004, Workshop on Terminology (2004)
15. Nichols, E., Bond, F., Tanaka, T., Sanae, F., Flickinger, D.: Multilingual ontology acquisition from multiple mrds. In: Proc. of 2nd Workshop on Ontology Learning and Population (OLP2), pp. 10–17 (2006)
16. Yeh, J.F., Wu, C.H., Chen, M.J., Yu, L.C.: Automated alignment and extraction of bilingual ontology for cross-language domain-specific applications. International Journal of Computational Linguistics & Chinese Language Processing 10(1), 35–52 (2005)
17. Ajani, G., Boella, G., Lesmo, L., Mazzei, A., Rossi, P.: Multilingual conceptual dictionaries based on ontologies. In: Proc. of V Legislative XML Workshop, pp. 1–14. European Press, Academic Publishing (June 2006)
18. Trojahn, C., Quaresma, P., Vieira, R.: Framework for multilingual ontology mapping. In: Proc. 6th Edition of the Language Resources and Evaluation Conference (LREC 2008). European Language Resources Association (ELRA) (2008)
19. Almeida, J.J., Simoes, A.: T2o recycling thesauri into a multilingual ontology. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odjik, J., Tapias, D. (eds.) Proc. of the 5th Int. Conf. on Language Resources and Evaluation (LREC 2006), pp. 1466–1471 (May 22-28, 2006)
20. Pazienza, M.T., Stellato, A.: An environment for semi-automatic annotation of ontological knowledge with linguistic content. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 11–14. Springer, Heidelberg (2006)
21. Espinoza, M., Perez, A.G., Mena, E.: Enriching an ontology with multilingual information. In: Proc. of 5th European Semantic Web Conference (ECSW 2008), pp. 333–347 (2008)
22. De Leenheer, P., de Moor, A., Meersman, R.: Context dependency management in ontology engineering: a formal approach. Journal on Data Semantics VIII, 26–56 (2007)
23. Magnini, B., Cavagli, G.: Integrating subject field codes into wordnet. In: Proceedings of LREC 2000, 2nd International Conference on Language Resources and Evaluation, pp. 1413–1418 (2000)
24. Stavrakas, Y., Gergatsoulis, M.: Multidimensional semistructured data: Representing context-dependent information on the web. In: Pidduck, A.B., Mylopoulos, J., Woo, C.C., Ozsu, M.T. (eds.) CAiSE 2002. LNCS, vol. 2348, pp. 183–199. Springer, Heidelberg (2002)
25. Ram, S., Park, J.: Semantic conflict resolution ontology (scrol): An ontology for detecting and resolving data and schema-level semantic conflicts. Transactions on Knowledge and Data Engineering (TKDE) 16(21), 189–202 (2004)
26. Marcu, D.: The rhetorical parsing of unrestricted texts: A surface-based approach. Computational Linguistics 26(3), 395–448 (2000)