From *Proceedings of the IFIP Conference on Data Semantics*, October 1997

Incorporating generalized quantifiers into description logic for representing data source contents

Steven Yi-cheng Tu and Stuart E. Madnick

Sloan WP #3998 CISL WP #98-01
January 1998

**The Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02142**

Incorporating generalized quantifiers into description logic for representing data source contents

S. Y. Tu and S. Madnick

Context Interchange Systems Laboratory
Sloan School of Management
Massachusetts Institute of Technology
Room E53-321, 30 Wadsworth Street, Cambridge, MA 02146 USA
Tel: 617/253-6671, Fax: 617/253-3321
{itu@mit.edu, smadnick@mit.edu}

Abstract

Systems for helping users to select data sources in an environment, such as the Internet, must be expressive enough to allow a variety of data sources to be formally represented. We build upon and extend the concept language, description logic (DL), to propose a novel representation system to achieve that goal. We point out that there are technical barriers within description logic limiting the types of data sources that can be represented. Specifically, we show that (1) DL is awkward in representing sufficient conditions, and (2) DL can describe properties of a concept itself only in the case of existential quantification. To be concrete, whereas it is easy in DL to say 'There are some objects in a concept C such that each of them holds the property P', it is awkward, if not impossible, to say 'All (Most, More-than-n, More-than-1/2) of those that hold the property P are in the concept C' or 'Most (More-than-n, More-than-1/2) of those that are in the concept C hold the property P'. These barriers cause us to extend DL with the notion of generalized quantifiers. We improve the previous results of generalized quantifiers to make them inter-operable with traditional logic. The proposed formalism integrates the nice features of generalized quantifiers into description logic, and hence achieves more expressive power than various representation systems based purely on description logic. It is also shown that our proposed language preserves those mathematical properties that traditional logic-based formalisms are known to hold.

Keywords

source description, description logic, generalized quantifier, logical transformation

1 INTRODUCTION

1.1 Motivation

Advances in computer networking technologies, such as the Internet, have tremendously widen the information access space. Nowadays it is technologically feasible that users can gain access, beyond their local reach, to many other remote data sources for topics ranging from business, education, to purely leisure. However, users in such an environment are usually confronted with the questions as to 'where are the data sources' and 'how different is each one'. Being able to effectively query certain data source presumes that availability of that particular data source and its distinction from other related ones have already been known to the users. This assumption, however, doesn't stand true in cases such as the Internet because of the large number and the rapid growth of data sources on the net. Conceivably, as more and more data sources are added, it will be increasingly formidable for users to perform source management by themselves. It is therefore a pragmatic and theoretical issue that theories be developed and system be designed to aid source selection.

One important issue about source selection concerns characterizing the contents of data sources and representing them in a formal fashion. The primary challenge of such aim lies upon the fact that data sources presently found on the Internet usually have contents that are very difficult to be described in a precise fashion. As a support, the following texts are extracted from a web page describing an online data source called *Company Capsules.*

'The capsules contain information on about 11,000 companies, including the largest and fastest-growing companies in the US and around the world.'

Capsules include:

- Every major publicly listed US company traded on the three major stock exchanges.
- More than 1,000 of America's largest private companies.
- More than 1,000 other private companies.
- More than 500 of the most important foreign companies.
- Most of the largest law firms and advertising agencies.

Although data sources about company financial information such as *Company Capsules* are numerous on Internet, it is uncommon that two data sources, both about companies, will have exactly the same contents. They differ either dramatically or subtly in terms of at least their **scopes** and **sizes**. For example, one data source called *CANADA/CD* contains Canadian companies, which differs from *Capsules* because of the geographical scopes being covered. Another data source *Worldscope* might also contain both US and international companies, but gives more foreign companies (e.g., most of the important foreign companies) than *Capsules*. In this context, if a user is interested in knowing the financial information about a particular US company, say 'General Motors', then it makes more sense for him/her to select *Company Capsules* as opposed to *CANADA/CD*, assuming other difference factors immaterial here. Similarly, if a user is looking for a foreign company, say 'Sumitomo Life Insurance Co.', then *Worldscope* should be given more precedence than *Company Capsules* because of the size distinction. It is therefore our observation that any practically useful system wanting to address the issue of source selection must take into accounts these scope and size differences, and reflect them in the descriptions of source contents.

Our research strategy is as follows:

1. The content of a data source is described by a target concept that corresponds to the populated instances stored in that data source. For example, *Company Capsules* will be described as a data source about the concept `Company`.

2. The target concept is characterized by a scope that is substantiated by a set of pre-defined characteristics. For example, `Company` is characterized by its geographical region (e.g., US or Canadian), its status (e.g., public or private), and its industry type (e.g., manufacturing or service), which all together defines a scope for this particular concept `Company`.
3. The target concept and associated characteristics are represented using an intensional assertion, which is accompanied by the size with respect to the scope (e.g., there are more than 500 of such companies).

Our research strategy is complementary to, not competing with, other approaches based upon the technique of keyword searching (e.g., search engines currently available on the net such as *AltaVista*) in two ways. Firstly, whereas we focus particularly on structured data sources, relational databases to be precise, most known search engines focus on unstructured or semi-structured web-page-like documents. Due to that structural difference, we are able to establish intensional assertions for relational database sources. Secondly, current search engines usually cannot directly access the contents of those relational data sources, but have to go through the interface of CGI. That is, source contents are dynamically generated as output of CGI programs. Our research strategy, regarding this point, has the advantage of making those hidden contents visible to a *source reasoner* that we are developing or to other keyword-based search engines.

We base upon **description logic (DL)** to propose a novel representation system to implement the above strategy. We point out that there are technical barriers within description logic limiting the types of data sources that can be represented. Specifically, we will show that (1) DL is awkward in representing sufficient conditions (Doyle and Patil, 1991) and (2) Though DL can describe the properties of objects in a concept in versatile ways, it can describe the properties of a concept itself only in the case of existential quantification (Quantz, 1992). For the first point, whereas it is easy in DL to represent 'All companies in data source R are US companies', it is difficult to say 'All US companies are in data source R'. For the second point, whereas it is easy in DL to represent the statement 'There are some courses in the concept `Graduate_Course` such that each of them must be taken by at least 3 (and at most 30) students', it is difficult to represent another statement 'There are 5 courses in the concept `Graduate_Course` with such property (e.g., taken by at least 3 students)'. These barriers cause us to extend DL with the notion of **generalized quantifiers**. We improve the previous results of generalized quantifiers to make them inter-operable with traditional first-order logic. As a result, the proposed formalism, as we will show, integrates the nice features of generalized quantifiers into description logic, and hence achieves more expressive power than various representation systems purely based on description logic. In a nutshell, we intend to advance the following theoretical arguments.

- Existing representational approaches to source selection only provide functionality at the level of relevance, but bear little to the notions of complete source and finer-grained characterization of relevant sources.
- Current systems are theoretically limited because of the types of quantifiers that can be asserted in their source description languages.
- Description Logic, underlying most current systems, can be enriched with results from generalized quantifier research, to enhance its usability for representing various source contents.
- Our proposed representation scheme, while powerful enough to accommodate a wide variety of source contents, enjoys the same mathematical properties that traditional logic-based languages are known to hold.

1.2 A running example

Consider a scenario that will be used throughout the rest of this paper. A user is interested in finding weather information regarding Taiwan's cities from the Internet. Specifically the user would like to have the request P_1 answered. Three data sources pertaining to city weather are listed below, where readers should note that Taipei, Koushung, Taichung, and Tainan are assumed to be all of Taiwan's major cities (Table 1).

<P_1>: What is yesterday's temperature for the city "Taipei"?

R1:		*R2:*		*R3:*	
CityName	*Yesterday Temp.*	*CityName*	*Yesterday Temp.*	*CityName*	*Yesterday Temp.*
Taipei	65 ‘F	Taipei	65 ‘F	Koushung	78 ’F
Koushung	78 ’F	Koushung	78 ‘F	Taichung	70 ’F
Taichung	70 ‘F	Taichung	70 ‘F	Tokyo	58 ‘F
Tainan	67 ‘F				
New York	41 ’F				

Table 1 A user request and three data sources.

The mathematical simplicity and uniformity of a relation, such as R_1-R_3, makes it possible to characterize its content intensionally. According to the Entity-Relationship model (Chen, 1976), a relation can be conceived as a collection of entities and attribute values where entities play the role of indicating existence of certain objects in the real world and attributes play a different role of providing descriptions for those entities. To concentrate discussions, we are only concerned about entity coverage but not attribute coverage in this paper. Every relation should have at least one key-identifying attribute, embodying the existence of certain objects, according to the unique entity integrity rule underlying the relational theory (Date, 1986). This property renders a representational alternative based upon the notion of intensionality, which is fundamentally different from the notion of extensionality; namely, enumerating the relation as a collection of extensional instances. For example, we can characterize R_1 intensionally below if that's what R_1's key attribute, `Cityname`, is supposed to be populated with. Likewise, R_2 and R_3 can also be characterized in the same way.

- R_1 contains {Major cities in Taiwan}, characterized intensionally.
- R_1 contains {Taipei, Koushung, Taichung, Tainan}, characterized extensionally.

Being that case, there is however one semantic gap pointed out by (Kent, 1978) regarding the two different notions of cities.

- The real set of existing major cities in Taiwan.
- The set of Taiwan's major cities currently represented in each data source.

To bridge this gap, we note that though data sources R_1-R_3 all contain information about Taiwan's major cities, the degree with respect to the real existing set that each individual source covers is unclear given the current descriptions. This point triggers the need to add coverage degree to the description of each individual data source, as the following literal statements demonstrate.

R_1: Contains all of the major cities in Taiwan.
R_2: Contains most of the major cities in Taiwan.
R_3: Contains some of the major cities in Asia.

The diversity of source content doesn't end up there. Let's consider other literal statements describing a range of data sources found on the Internet, of which most are real data sources.

R_4: Contains more than 3 major cities in Taiwan.

R_5: Contains more than 1/2 of the major cities in Taiwan.

R_6: Contains 2 major cities for every country located in Asia.

The above characterization for R_1-R_6 represent the case that the entities contained in each source are quantified toward the real set of existing major cities in Taiwan. They illustrate a sufficient condition for entities that might appear in each data source. Unsurprisingly, data sources that can only be asserted in the form of necessary condition have the equal chance to occur as those in sufficient conditions. R_7-R_{12} illustrate this point. It is useful to note that, unlike sufficient conditions, necessary conditions are quantified toward the set of entities within the data source itself, not the existing entities (e.g., major cities in Taiwan) in the real world. Readers are invited to pay special attention to R_6 and R_{12} because they constitute the two most challenging cases that highlight the superiority of our proposed language.

R_7: All of the contained are major cities in Taiwan.

R_8: Most of the contained are major cities in Taiwan.

R_9: Some of the contained are major cities in Taiwan.

R_{10}: More than 3 of the contained are major cities in Taiwan.

R_{11}: More than 1/2 of the contained are major cities in Taiwan.

R_{12}: Most of the contained are major cities in Asia, most of which are in Taiwan.

1.3 Importance of quantifiers

In the case that there is a universally quantified data source about major cities in Taiwan like R_1, answering P_1 can be performed in a very effective fashion. To see that, let $R_1(x)$ be a predicate that is assigned true if x is in R_1, and U, the domain x ranges over, be the real set of Taiwan's major cities, then the following inference rule deduces that the particular city 'Taipei' can always be found in R_1. In other words, R_1 is guaranteed to be able to answer P_1. Such a data source is said to have completeness knowledge (Etzioni, Golden, and Weld, 1994, Motro, 1989) with respect to a scope.

{U = the existing set of Taiwan cities, $x \in U$, $\forall x R_1(x)$} $\mapsto$ R_1("Taipei")

or {U = the existing set of things, $x \in U$, $\forall x$ (Taiwan_city(x)$\Rightarrow$ R_1(x))} $\mapsto$ R_1("Taipei")[1]

Data sources with completeness knowledge are undeniably effective for answering users' queries in the form of intensional reasoning. However, as other data sources suggest, establishing completeness knowledge with respect to the real world is not common. In most cases, data sources, as R_2 exemplifies, only contain incomplete knowledge. There are problems when asserting incomplete data sources with quantifiers other than the traditional universal and existential quantifiers if the rudimentary semantics of quantifiers are not well handled. Consider R_2 as an example.

U = the existing set of things, $x \in U$, Most x (Taiwan_city(x) $\Rightarrow$ R_2(x))

The above characterization of R_2 is a wrong one. In fact as argued by (Sher, 1991), any generalized quantifiers (e.g., most, more-than-3, more-than-1/2, less-than-1/3 and etc.) will fail when they are intended to operate with logical conditional. The explanation in the context of this particular assertion is as follows. Suppose that the universe U is $\{\vartheta, \lambda, \mu, \nu, o, \rho, \sigma, \zeta, \psi\}$, and the corresponding truth table is as listed in Table 2. It is clear that Most x(Taiwan_city(X) $\Rightarrow$ R_2(X)) should be true because most elements out of the universe (e.g., six out of eight elements)

[1] In this paper, $\Rightarrow$ denotes logical conditional and $\mapsto$ denotes logical implication (deduction).

turn out to be true for the whole assertion. But obviously it should be false because there are more elements (e.g., ρ and σ) that are in Taiwan_city(x) but not in $R_2(X)$, than those are (i.e., ζ). There is a logical fallacy here. In other words, the whole assertion is true not because most of Taiwan's major cities are in R_2, but because most things in the universe are not Taiwan's major cities. The impact is that although in reality only R_2 contains most of the Taiwan's major cities, this assertion is, however, mistakenly true for R_3-R_6. Note however that the assertion for R_1; namely, $\forall x(\text{Taiwan\_city}(x) \Rightarrow R_1(x))$, doesn't involve this logical problem, mainly because of the particularity of the universal quantifier. This logical fallacy being present, we have a situation where for those data sources that can only be asserted by generalized quantifiers, the semantics of logical conditional must be re-examined. We will address this issue in Section 3 where we follow a research line that has been rather successful in providing a coherent explanation of semantics of generalized quantifiers through the set-theoretical view.

	Taiwan_city(x)	*R2(x)*	*Most x (Taiwan_city(x) ⇒ R2(x))*
ϑ	False	-	True
λ	False	-	True
μ	False	-	True
ν	False	-	True
ρ	True	False	False
σ	True	False	False
ζ	True	True	True
ψ	False	-	True

Table 2 A truth table involving a generalized quantifier and logical conditional.

1.4 Paper organization

The rest of this paper is organized as follows. In Section 2, we review and revise some important tools required to build our representation language. It covers topics from description logic, quantification function, and generalized quantifiers. Section 3 is devoted to finding a means to incorporate generalized quantifiers into DL to come up with a powerful, yet semantically consistent, language. The syntax and semantics of the assertion language will be presented there. We use the source examples R_1-R_{12} to show that the language is rather expressive relative to other systems purely based on description logic. Also contained in this section are a set of mathematical properties that are shown to hold for our language. In Section 4, we discuss the applications and potential implications of our proposed language and lay out a set of factors affecting its computational behaviors.

2 MECHANISM

Three tools constitute the key elements of our representation system. They are description logic for representing scope differences, generalized quantifiers for representing size differences, and quantification functions for uniformly explicating the semantics of quantifiers. These are presented in the context of the aforementioned source examples.

2.1 Description Logic

Description Logic (DL), also known as terminological logic, is strongly related to frame-like languages and has been used as a formalism underlying various knowledge representation systems, such as ARGON (Patel-Schnider, Brachman, and Levesque, 1984), KL-ONE (Brachman and Schmolze, 1985), KRYPTON (Brachman, Fikes, and Levesque, 1983), and LOOM (MacGregor and Bates, 1987). Recently there have been efforts applying DL in the database field (Anwar, Beck, and Navathe, 1992, Borgida and Brachman, 1993, Beneventano, Bergamanschi, and Lordi, 1994, Borgida, Brachman, and McGuinness, 1989, Beck, Gala, and Navathe, 1989, Borgida, 1995, Bresciani, 1995, Bresciani, 1996, Bergmschi and Sartori, 1992, Devanbu, 1993, Kessel, Rousselot, Schlick, and Stern, 1995), mainly because of the capabilities that are considered fundamental in semantic data modeling such as structural description and taxonomic classification (Hull and King, 1987) are inherently equipped within DL. Structural constructors pivotal within DL are *concept*, *individual*, and *role* where an individual represents a single object, a concept is a collection of individuals, and a role is a relation associating a pair of concepts or individuals. A new concept can be defined by conjoining concepts that are already defined or by adding *role restriction* and *numeric quantifier restriction* to a defined concept (Calvanese, Lenzerini, and Nardi, 1992, Hollunder and Baader, 1994, MacGregor, 1994). DL's clean syntax (e.g., dot operator and set connectives) comprises of a set of syntactic constructors for complex DL expressions to be formulated, which basically represent classification and subsumption relations among concepts

Systems built on DL usually have two components called *T-box* (Terminological box) and *A-Box* (Assertion box) (Brachman, Fikes, and Levesque, 1983, Giacomo and Lenzerini, 1996). The former box functions like noun-phrase terms that can be referenced while interacting with the system, and the latter functions like propositional sentences that relate the terminological objects in T-box. This separation enormously influences subsequent research on system integration and knowledge representation (Arens, Chee, Hsu, and Knoblock 1993, Knoblock, Yigal, and Hsu, 1994, Levy, Srivastava, and Kirt, 1995, Levy, Rajarman, and Ordille, 1996, Mays, Tyler, McCuire, and Schlossberg, 1987). In our approach, terms appearing in T-box, including individuals, concepts, and roles, correspond respectively to the instance names, target concepts, and concept characteristics. Assertions stored in A-box are to describe the content of each data source by referring to the terms in T-box. Each data source is linked to a target concept, which is characterized by a set of roles and corresponding concepts in T-Box (see Table 3).

The instance name 'Taipei' appearing in P_1 can be modeled as an individual of the concept, `City`. A concept schema for `City`, which echoes the notion of source characteristics, can be written as E_1, where `THING`, denoted as $\perp$, is the top primitive concept built into the system. The target concept `City` is characterized by two primitive concepts, `Region` and `Status` through respectively the roles `Located` and `Legal`. The data source R_0 is linked to the concept schema as in E_2[2] where `Located:Taiwan` corresponds to the `Filler` operations in DL. Note that the two universal quantifiers appearing in E_1 represents role restriction, which in this case restricts the values related to the individuals in `City` through the roles `Located` and `Legal` be drawn from only `Region` and `Status` respectively.

[2] In this paper, $\sqcap$ denotes the concept conjunction operation conventionally used in DL, and $\cap$ denotes the set-intersection connective.

T-box:

Taiwan $\in$ Region Asia $\in$ Region World $\in$ Region

Major $\in$ Status Capital $\in$ Status Non-capital $\in$ Status Minor $\in$ Status

Town $\in$ Status Legal_city $\in$ Status City_town $\in$ Status

E_1: City $\sqsubset$ $\perp$ $\sqcap$ $\forall$Located.Region $\sqcap$ $\forall$Legal.Status

A-box:

E_2: R_0 $\sqsubset$ City $\sqcap$ Located:Taiwan $\sqcap$ Legal:Major

Table 3 An example of T-box and A-box.

From first-order logic point of view, a concept in DL is a unary predicate and a role is a binary predicate drawing domains from two concepts. Role restriction and subsumption relations are defined on the model-theoretical semantics where a concept C_1 subsumes another concept C_2 just in case the extension of C_2 is a subset, not necessarily proper, of the extension of C_1 for unary predicates (Borgida, 1995). Role subsumption is defined the same way for binary predicates. The following Table 4 (Patel-Schneider and Swartout, 1993) summarizes the syntax and model-theoretical semantics of DL's constructs relevant to this paper.

Syntax input	*Syntax abstract*	*Semantics*
TOP	$\perp$	Δ^I
(Define-primitive-concept CN Concept)	$CN \sqsubset C$	$CN^I \subseteq C^I$
(Define-concept)	$CN \sqsubset\sqsupset C$	$CN^I = C^I$
(And Concept$_1$ Concept$_2$....Concept$_n$)	$C_1 \sqcap C_2 \sqcap \ldots \sqcap C_n$	$C_1^I \cap C_2^I \cap \ldots \cap C_n^I$
(Instance Individual Concept)	$IN \in C$	$IN^I \in C^I$
(All Role Concept)	$\forall$ R.C	$\{d \in \Delta^I \mid R^I(d) \subseteq C^I\}$
(Filler Role Individual)	R:i	$\{d \in \Delta^I \mid R^I(d) \subseteq i^I\}$
(Subset Role$_1$ Role$_2$)	$R_1 \sqsubset R_2$	$R_1^I \subseteq R_2^I$
(At-least n Role Concept)	$\geq$ n R.C	$\{d \in \Delta^I \mid \lvert R^I(d) \cap C^I \rvert \geq n\}$
(Inverse Role)	R^{-1}	$R^I \cap (\Delta^I \times \Delta^I)$
(Not Concept)	$\neg C$	$\Delta^I - C^I$

Table 4 A summary of DL's syntax and semantics

Since the semantics of DL is based upon the model-theoretical view, subsumption in DL therefore has a strong synergy with the idea of logical conditional. In other words, the following characterizations between the two concepts should all be considered equivalent (Baader, 1996).

<DL perspective - subsumption>: $C_1 \sqsubset C_2$

<Set-theoretical perspective - containment>: $C_1^I \subseteq C_2^I$

<First-order logic perspective - conditional>: $\forall$ x $(C_1(x) \Rightarrow C_2(x))$

Following the above analysis, we can actually re-write E_1 and E_2 in a more logic-like form (Baader, 1996). For instance, let U be the universe from which all individuals (i.e., Δ^I) are drawn, E_3 and E_4 rewrite E_1 and E_2 respectively into the form of first-order logic. Logically, E_3 and E_4 should be evaluated time-invariantly true in the system if they are meant be equivalent with E_1 and E_2.

E_3: $\forall$x City(x) $\Rightarrow$ (THING(x) $\wedge$ ($\forall$y Located(x, y) $\Rightarrow$ Region(y)) $\wedge$ ($\forall$z Legal(x, z) $\Rightarrow$ Status(z))

E_4: $\forall$x R_0(x) $\Rightarrow$ (City (x) $\wedge$ Located(x, Taiwan) $\wedge$ Legal(x, Major))

Concerning the role `Located` in E_4, it says that the city x is located in the region Taiwan. Supposing that every city located in Taiwan is also located in Asia, which is in actuality true, then it is valid to replace the predicate Located(x,Taiwan) with Located(x,Asia) and the whole assertion still remains true. This point actually has some impacts on the problem of answering queries because if a request is asking for cities located in Taiwan, then a data source containing Asian cities should also be considered as a plausible data source for answering that request. In light of this relationship, those individuals within a concept with this property are organized into a hierarchy where the constraint of *inverse role subsumption* (Calvanese, Lenzerini, and Nardi, 1992) should hold among levels of individuals. For example, all cities in Taiwan are in Asia, and all cities in Asia are in the world all the way but not vice verse (see Figure 1). We call such relationship between two individuals a *role individual subsumption* relationship, which is denoted as $\uparrow$.

<Role individual subsumption>	$\uparrow$(Taiwan, Asia), $\uparrow$(Asia, World) $\uparrow$(Capital, Major), $\uparrow$(Non-Capital, Major), $\uparrow$(Major, Legal_City), $\uparrow$(Minor, Legal_City), $\uparrow$(Legal_City, City_town), $\uparrow$(Town, City_town)

<DL >	Located(x, Taiwan) $\sqsubset$ Located(x, Asia) $\sqsubset$ Located(x, World)
	Legal(x, Capital) $\sqsubset$ Legal(x, Major) $\sqsubset$ Legal(x, Legal_city) $\sqsubset$ Legal(x, City_town)
	Legal(x, Non-capital) $\sqsubset$ Legal(x, Major), Legal(x, Minor) $\sqsubset$ Legal(x, Legal_city)
	Legal(x, Town) $\sqsubset$ Legal(x, City_town)

<Set>	Located^{-1}(Taiwan) $\subseteq$ Located^{-1}(Asia) $\subseteq$ Located^{-1}(World)
<Logic >	$\forall$x Located(x, Taiwan) $\Rightarrow$ Located(x, Asia) $\Rightarrow$ Located(x, World)

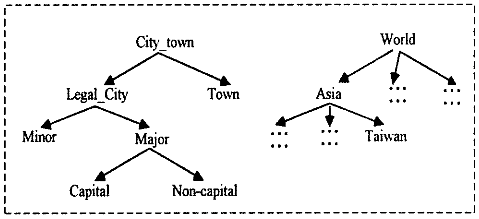

Figure 1 Levels of individuals in concepts Region and Status.

2.2 Quantification function

Universal and existential quantifiers, though typical in traditional logic, are atypical in human language. A lot of knowledge that prevail in worldly usage can only be stated in an imprecise form (Westerstahl, 1989), as our previous source examples indicate. Accordingly, allowing generalized quantifiers to be used in source assertions will be of great practical value. To accomplish that goal begs for a fundamental question needed to be investigated as to the nature of quantifiers. In other words, what is a quantifier? Historical studies in mathematical logic provided a clue to unfold this question. According to an influential work about quantifiers (Frege, 1968), Frege contended that a quantifier Q_τ is a second-level object that takes the first-level logical predicate object as input and returns either truth or falsity, where the truth value depends only on the size of the extension satisfying the first-level predicate. Namely, every formula $Q_\tau(\Phi)$ is associated with a quantification function f_τ: $P(U) \rightarrow \{\text{True, False}\}$, where P(U) is the power set of U. The function f_τ is assigned true if the model Γ of the logical formula Φ satisfies the cardinality relation specified by Q_τ, and is assigned false otherwise. In other words, the logic formula $Q_\tau(\Phi)$ is true if and only if f_τ $(\Gamma) \rightarrow$ True. Let's see how the conventional universal and existential quantifiers can be defined on this basis. Again U denotes the universe of discourse.

$f_\forall(\Gamma) =$ True if $|U - \Gamma| = 0$;otherwise False

$f_\exists(\Gamma) =$ True if $|\Gamma| > 0$;otherwise False

A quantification function can also be encoded as a pair (γ, α) such that $\gamma+\beta = \alpha$, where α is the size of the universe, γ is the size of the model Γ for Φ, and β is the size of the complement of Γ. The beauty of Frege's characterization of quantifiers is that it provides a uniform treatment of quantifiers using only cardinal numbers. Mostowski, following the same course, laid out a mathematical foundation for generalized quantifiers based on quantification functions in his seminal paper (Mostowski, 1957). Some generalized quantifiers that frequently come across in human's language follow, out of which we will assume that the literal term "Most" is interpreted as "more than 2/3" of the set in question.

$f_{\text{Most}}(\Gamma) =$	True	if	$\gamma \geq 2/3\ \alpha$	;otherwise False
$f_{\geq 3}(\Gamma) =$	True	if	$\gamma \geq 3$	;otherwise False
$f_{=2}(\Gamma) =$	True	if	$\gamma = 2$	;otherwise False
$f_{\text{more-than-1/2}}(\Gamma) =$	True	if	$\gamma > 1/2\ \alpha$ or $\gamma > (\alpha - \gamma)$	;otherwise False

Given the above quantification functions available, we can describe the contents of R_2-R_6 using quantification functions. We can see that those data sources such as R_2-R_6, which can only be existentially quantified in traditional logic, are now differentiated in a finer manner through their associated quantification functions.

R_1: quantifier $Q_{\forall}$ with the quantification function $f_{\forall}$

R_2: generalized quantifier Q_{Most} with the quantification function f_{Most}

R_3: quantifier $Q_{\exists}$ with the quantification function $f_{\exists}$

R_4: generalized quantifier $Q_{\geq 3}$ with the quantification function $f_{\geq 3}$

R_5: generalized quantifier $Q_{\text{More-than-1/2}}$ with the quantification function $f_{\text{More-than-1/2}}$

R_6: quantifier $Q_{\forall}$ and generalized quantifier $Q_{=2}$ with the quantification functions $f_{\forall}$ and $f_{=2}$

2.3 Generalized quantifiers

Mostowski's original and rigorous work inspired immense intellectual interests on the logical and algebraic properties of generalized quantifiers, and since then has spawned a line of elegant theory for researching the nature of generalized quantifiers (Barwise and Cooper, 1981, Benthem, 1982, Sher, 1991). One important result that is of our great interest here is that generalized quantifiers fail for logical conditional (i.e., ⇒) as we mentioned in Section 1.3 about the "Most" quantifier. Barwise (Barwise and Cooper, 1981), aiming at that problem, proposed a rudimentarily different perspective in studying generalized quantifiers. He dismissed the perspective that a quantifier is a logical symbol and switched to view quantifiers as noun phrases. In brevity, Barwise considered a traditional quantifier as a determiner, and it must be supplemented with a set expression to compose an operative quantifier (see Figure 2). Therefore, the meanings of determiners like 'Most' or 'More than 3' are open unless the sets to be modified by these determiners are given.

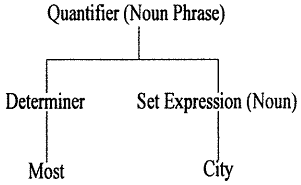

Figure 2 Barwise's conception of quantifiers.

The following syntax and semantics illuminate Barwise's stance, where again U denotes the universe of discourse.

All (A's, B's)	$\{A \subseteq U \mid \lvert A \cap B\rvert = \lvert A\rvert\}$
Most (A's, B's)	$\{A \subseteq U \mid \lvert A \cap B\rvert > \lvert A - B\rvert\}$
Some (A's, B's)	$\{A \subseteq U \mid \lvert A \cap B\rvert > 0\}$
>n (A's, B's)	$\{A \subseteq U \mid \lvert A \cap B\rvert > n\}$

Generalized quantifiers, though potentially promising to the database management area, have been surprisingly under-explored. Recently, there are two works related to this area, (Hsu and Parker, 1995) and (Quantz, 1992). The former paper made an attempt to demonstrate the value of adding generalized quantifiers into traditional terminological logic in the domain of biology (i.e., bound anaphora), in particular on the issue of referential representation for object disambiguation (Quantz, Schmitz, and Kussner, 1994). The latter paper showed the usefulness of generalized quantifiers in query formulation and proposed an approach to extending SQL with the ability to process generalized quantifiers.

Barwise's conception of generalized quantifiers' for resolving the case of logical conditional allow us to write down the assertions of those data sources for which generalized quantifiers are needed. However, readers can note that the contents of R_6 and R_{12} still cannot be uttered within Barwise's framework. The reason can be attributed to, as Sher criticized in (Sher, 1991), Barwise's dismissal of quantifiers as logical symbols. According to Mostowski, a quantifier should be formula-building and enable us to construct propositions. Namely, syntactically quantifiers must allow us to bind free variables appearing in their attached formulas to generate more complex formulas. In that regard, Barwise's forsaking generalized quantifiers as logical symbols makes it difficult to syntactically formulate propositions that involve first-order logic variable (e.g., R_6) or involve more than two formulas both containing generalized quantifiers (e.g., R_{12}). It is thus fair to say that Barwise's work, though useful in forwarding a consistent approach to handling logical conditional with generalized quantifiers, actually loses some compositional expressiveness. This motivates us to seek a language that still maintains the spirit of Barwise's conception of generalized quantifiers and yet is inter-operable with first-order logic formula so that source contents such as R_6 and R_{12} can be expressed.

R_1:	$Q_{\forall}$ ({Major cities in Taiwan}, R_1)
R_2:	Q_{Most} ({Major cities in Taiwan}, R_2)
R_3:	$Q_{\exists}$ ({Major cities in Taiwan}, R_3)
R_4:	$Q_{\geq 3}$ ({Major cities in Taiwan}, R_4)
R_5:	$Q_{More\text{-}than\text{-}1/2}$ ({Major cities in Taiwan}, R_5)
R_6:	???
R_7:	$Q_{\forall}$ (R_7, {Major cities in Taiwan})
R_8:	Q_{Most} (R_8, {Major cities in Taiwan})
R_9:	$Q_{\exists}$ (R_9, {Major cities in Taiwan})
R_{10}:	$Q_{\geq 3}$ (R_{10}, {Major cities in Taiwan})
R_{11}:	$Q_{More\text{-}than\text{-}1/2}$ (R_{11}, {Major cities in Taiwan})
R_{12}:	???

3 EMBED GENERALIZED QUANTIFIERS IN DL

In spite of our heavy drawing upon description logic and Barwise's program of generalized quantifiers, there are however difficulties in using them to represent certain types of data sources. For description logic, one difficulty rests upon the observation that DL is better suited for representing integrity knowledge, but not for representing completeness knowledge. The other difficulty is, as pointed out by (Quantz, 1992), description logic can only assert properties of objects in a set, but not properties of the set itself as an object. In our city example, whereas it is easy to say 'All cities in the data source R are Taiwan's major cities', it is awkward to say 'All (Most, More-than-n, More-than-1/2) of those Taiwan's major cities are in R' or 'Most (More-than-n, More-than-1/2) of those that are in R are Taiwan's major cities'. Barwise's generalized quantifiers arises to remedy this deficiency. Yet his limited view excludes the possibility of having traditional first-order logic variables as part of the formulas that generalized quantifiers are associated with. Targeting at those difficulties, we accordingly propose our solution in more formal detail in this section.

3.1 Limit of prior work

One problem about DL concerns its awkwardness in representing sufficient conditions, and due to that reason completeness knowledge is difficult to be asserted. Let's examine the problems when trying to use DL to assert R_1's content, which has complete knowledge about major cities in Taiwan. Three possible alternatives are shown in E_5-E_7, which are all unfeasible due to the following explanations. E_5 is not describing what R_1 contains because it only says that all those in R1 are Taiwan's major cities, which is essentially a kind of integrity knowledge as opposed to completeness knowledge. It is the former kind of knowledge because the semantics of `Define-primitive-concept` enforce that the extension of the left-hand side is a subset of the extension of the right-hand side. It is therefore tempting to choose E_6 instead using the DL primitive `Define-concept`, since now the set of Taiwan's major cities are contained in R_1. Yet E_6 is still an awkward characterization because it is an overstatement in the sense that the primitive "[ ]" also implies necessary condition, which is untrue because R_1 also covers some US cities. The final attempt may be to put `City` at the left-hand side as E_7. This alternative still doesn't work because in this case R_1 must have been defined somewhere by `City`, which inevitable will cause a definitional cycle (Baader, 1990). In other words, while "$\sqsubset$" represents necessary condition and "[ ]" represents equivalency, it is unclear how sufficient condition alone can be stated in a natural way in DL. As a consequence, it is awkward, if not impossible, using DL to express complete data sources. In that regard, a logical alternative, as E_8 shows, outperforms DL and allows completeness knowledge to be represented.

E_5: $R_1 \sqsubset$ City $\sqcap$ Located: Taiwan $\sqcap$ Legal: Major

E_6: R_1 [ ] City $\sqcap$ Located: Taiwan $\sqcap$ Legal: Major

E_7: City $\sqcap$ Located:Taiwan $\sqcap$ Legal: Major $\sqsubset R_1$

E_8: $\forall x\ ((\mathrm{City}\ (x) \wedge \mathrm{Located}(x, \mathrm{Taiwan}) \wedge \mathrm{Legal}(x, \mathrm{Major})) \Rightarrow R_0(x))$

The other problem about DL is its inability to represent other quantifiers besides the existential quantifier for describing properties at the set level. Although numeric quantifier restrictions are equipped within DL, which look like generalized quantifiers, they however can only be applied to the objects of a concept through *role restriction* (Baader, 1996). Consider the following E_9 as an example, which says 'all graduate courses must be taken by at least 3 students'. E_{10} shows the logical representation for E_9. That statement is essentially distinct from another statement 'there are more than 3 of such courses' as stated by E_{11} using generalized quantifiers. Accordingly, we can note that numeric quantifier restriction in DL can be used to specify properties only for the existential case at the set level, and only for the universal case at the object level. In other words, DL basically, in this paper's context, allows us to say 'There are some cities with respect to the real world in a data source R such that each of which holds the property P (e.g., cities located in Taiwan)'. Yet it is awkward, if not impossible, to say 'All (Most, More-than-n, More-than-1/2) of those that hold the property P are in a data source R' or 'Most (More-than-n, More-than-1/2) of those that are in a data source R hold the property P'.

E_9: Graduate_course $\sqsubset$ Course $\sqcap$ ≥ 3 Taken.Student

E_{10}: $\forall$ x Graduate_course(x) $\Rightarrow$ (Course(x) $\wedge$ Student(s) $\wedge$ ≥ 3 Taken(x, s))

E_{11}: $Q_{\geq 3}$ (Graduate_course(x), Taken(x, s))

The discussions so far suggest that there are representational benefits to be gained if generalized quantifiers are incorporated into DL. Yet means must be sought to avoid Barwise's syntactic limit such that R_6 involving first-order logic variable and R_{12} involving more than one applications of generalized quantifiers can also be expressed.

3.2 An expressive language

In general, Barwise considered that for any quantifier Q_τ, the formula Q_τ(A's, B's) is a set-theoretical relation between A and B, where τ explicates that relation through its quantification function f_τ. We can embed this notion into description logic by defining a new operator, $\rightarrow_{Q_\tau}$, such that $A(\overset{=}{x}\,;\, x_1 \sim x_i) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\, y_1 \sim y_j)$ is true if and only if $Q_\tau(A(\overset{=}{x}), B(\overset{=}{x}))$ is true with the following restrictions.

- $A(\overset{=}{x}\,;\, x_1 \sim x_i)$ is a conjunctive DL expression involving i+1 variables.
- $B(\overset{=}{x}\,;\, y_1 \sim y_j)$ is a conjunctive DL expression involving j+1 variables.
- $x_1 \sim x_i$, $y_1 \sim y_j$ must be either constants or bound variables by either $\exists$ or $\forall$ quantifier.
- $\overset{=}{x}$ is a free variable.
- There cannot be generalized quantifier appearing within A or B.

The semantics of the new operator is defined following the semantics of the generalized quantifier Q_τ along with its quantification function f_τ encoded as the pair (γ, α). Literally, we can translate the sentence, $A(\overset{=}{x}\,;\, x_1 \sim x_i) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\, y_1 \sim y_j)$, into the statement 'If the premise predicate $A(\overset{=}{x}\,;\, x_1 \sim x_i)$ is true, then there are γ $\overset{=}{x}$'s in $A(\overset{=}{x}\,;\, x_1 \sim x_i)$ out of $\alpha\,\overset{=}{x}$'s that also satisfy the consequence predicate $B(\overset{=}{x}\,;\, y_1 \sim y_j)$, where again α is the cardinality of the model satisfying $A(\overset{=}{x}\,;\, x_1 \sim x_i)$, and γ is the cardinality of the model satisfying both A and B'. We can note that the mapping between the operator $\rightarrow_{Q_\tau}$ and Barwise's form of generalized quantifiers is always valid because the only free variable in A and B is $\overset{=}{x}$, which guarantees that the generalized quantifier Q_τ is applied to two sets of the same sort. In this case, the two sets are composed of only unary tuples. Given that semantic correspondence, our approach is advantageous compared with Barwise's approach in two ways. Firstly, first-order variables quantified by traditional universal and existential quantifiers are now allowed within the premise and consequence predicates that build up the formula involving generalized quantifiers. Secondly, we can now consider $A(\overset{=}{x}\,;\, x_1 \sim x_i) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\, y_1 \sim y_j)$ as a regular first-order atomic formula, which can be connected via traditional logical connectives, such as conjunction, disjunction, and negation[3], to construct more complex formulas. The reason for the second advantage is that the formula $A(\overset{=}{x}\,;\, x_1 \sim x_i) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\, y_1 \sim y_j)$, although behaving in the sense of set relation from within the parentheses, is of pure logical sense from the view outside the parentheses (i.e., the interpretation space of the whole sentence is still {truth, falsity}). We now

[3] In this paper, $\wedge$ denotes logical conjunction, $\vee$ denotes logical disjunction, and $\neg$ denotes logical negation.

show how this approach can express the contents of R_1-R_{12}. Again, readers can pay special attention to R_6 and R_{12} because they support respectively the first and second advantages.

R_1: $City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major) \rightarrow_{Q_\forall} R_1(x)$

R_2: $City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major) \rightarrow_{Q_{Most}} R_2(x)$

R_3: $City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major) \rightarrow_{Q_\exists} R_3(x)$

R_4: $City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major) \rightarrow_{Q_{\geq 3}} R_4(x)$

R_5: $City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major) \rightarrow_{Q_{More-than-1/2}} R_5(x)$

R_6: $\forall y\ Region(y) \wedge \uparrow(y, Asia) \Rightarrow (City(x) \wedge Located(x, y) \wedge (Legal(x, Major) \rightarrow_{Q_{=2}} R_6(x))$

R_7: $R_7(x) \rightarrow_{Q_\forall} City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major)$

R_8: $R_8(x) \rightarrow_{Q_{Most}} City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major)$

R_9: $R_9(x) \rightarrow_{Q_\exists} City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major)$

R_{10}: $R_{10}(x) \rightarrow_{Q_{\geq 3}} City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major)$

R_{11}: $R_{11}(x) \rightarrow_{Q_{More-than-1/2}} City(x) \wedge Located(x, Taiwan) \wedge Legal(x, Major)$

R_{12}: $(R_{12}(x) \rightarrow_{Q_{Most}} City(x) \wedge Located(x, Asia)) \wedge (R_{12}(x) \wedge City(x) \wedge Located(x, Asia) \rightarrow_{Q_{Most}}$ $Located(x, Taiwan))$

We now define more formally, through the notions of well-formed formula and logical satisfaction, the syntax and semantics of our language combining both traditional universal and existential quantifiers and generalized quantifiers.

<Definition 1>: An assertion ε is a **well-formed formula (wff)** in our language Ω provided that:

(a) If ε is in the set of first-order wffs expressed in DL form, then ε is in Ω.

(b) If ε is $A(\overline{\overline{x}}; x_1 \sim x_i) \rightarrow_{Q_\tau} B(\overline{\overline{x}}; y_1 \sim y_j)$, then ε is in Ω.

(c) If ε1 and ε2 are in Ω, then so are $\varepsilon 1 \vee \varepsilon 2$, $\varepsilon 1 \wedge \varepsilon 2$, $\varepsilon 1 \Rightarrow \varepsilon 2$, $\varepsilon 1 \Leftrightarrow \varepsilon 2$, and $\neg \varepsilon 1$.

(d) If ε doesn't satisfy any of the above, then ε is not in Ω.

<Definition 2>: A wff ε in our language Ω is said to be true with the model Γ provided that:

(a) All first-logic formulas are true under Γ.

(b) All generalized-quantifier formulas of the form $A(\overline{\overline{x}}; x_1 \sim x_i) \rightarrow_{Q_\tau} B(\overline{\overline{x}}; y_1 \sim y_j)$ are true under Γ.

(c) Every universal quantification function is true: $f_\forall(\Gamma) \rightarrow$ true.

(d) Every existential quantification function is true: $f_\exists(\Gamma) \rightarrow$ true.

(e) Every generalized quantification function is true: $f_\tau(\Gamma) \rightarrow$ true.

(f) ε is not true if any of the above is not true.

Definition 1 and 2 concern respectively the syntax and semantics of our proposed language. It is clear from the definitions that our language differs from other DL-based languages primarily in the addition of the term $A(\overline{\overline{x}}; x_1 \sim x_i) \rightarrow_{Q_\tau} B(\overline{\overline{x}}; y_1 \sim y_j)$, which essentially achieves the capability of generalized quantifiers we intend to capture in our language. The proposed language being defined as such, there are two questions that we need to investigate. The first question pertains to the syntactic transformation of formulas, and the second question pertains to the deductibility of a set of formulas. Both questions are fundamental to the studying of logical languages (Enderson, 1971), and are discussed in the format of theorems organized into three related groups in the next sub-section.

3.3 Properties

The first group of theorems[4] consider two aspects, one being the relations of equivalency and implication between traditional quantifiers and their corresponding representations using our formalism, and the other being the transformability of the formulas in our language into *Prenex Normal Form* (i.e., all first-order quantifier symbols are left to other symbols). The significance of the following theorems can be illustrated from Theorem 1.3 and 1.4, which say that between the 'maximal' (i.e., universal quantifier) and the 'minimal' quantifier (i.e., existential quantifier), there are actually other non-traditional (i.e., generalized) quantifiers between these two extreme cases. In the context of this paper, it means that we can have more diverse ways of describing data source contents besides only being allowed to say they are either universally or existentially quantified. For presentation simplicity, we abbreviate $A(\overset{=}{x}\,;x_1 \sim x_i) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;y_1 \sim y_j)$ as $A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x})$ in the sequel wherever there is no notional confusion.

<Theorem 1: Equivalency, implication>: The following properties hold for the language Ω.

<1.1> $\forall\overset{=}{x}\, A(\overset{=}{x}) \Rightarrow B(\overset{=}{x}) \mapsto\!\leftarrow\!\!\dashv A(\overset{=}{x}) \rightarrow_{Q_\forall} B(\overset{=}{x})$

<1.2> $\exists\overset{=}{x}\, A(\overset{=}{x}) \Rightarrow B(\overset{=}{x}) \mapsto\!\leftarrow\!\!\dashv A(\overset{=}{x}) \rightarrow_{Q_\exists} B(\overset{=}{x})$

<1.3> $A(\overset{=}{x}) \rightarrow_{Q_\forall} B(\overset{=}{x}) \mapsto A(\overset{=}{x}) \rightarrow_{Q_\exists} B(\overset{=}{x})$

<1.4> $A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x})) \mapsto A(\overset{=}{x}) \rightarrow_{Q_\exists} B(\overset{=}{x})$ if τ is not the generalized quantifier "None"

<Theorem 2: Prenex Normal Form>: The following properties hold for the language Ω.

<2.1> $\neg\forall\overset{=}{x}\; A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x}) \mapsto\!\leftarrow\!\!\dashv \exists\overset{=}{x}\; \neg(A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x}))$

<2.2> $\neg\exists\overset{=}{x}\; A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x}) \mapsto\!\leftarrow\!\!\dashv \forall\overset{=}{x}\; \neg(A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x}))$

<2.3> $(C(x) \Rightarrow (\forall\overset{=}{x}\; A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x}))) \mapsto\!\leftarrow\!\!\dashv \forall\overset{=}{x}\; (C(x) \Rightarrow (A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x})))$

<2.4> $(C(x) \Rightarrow (\exists\overset{=}{x}\; A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x}))) \mapsto\!\leftarrow\!\!\dashv \exists\overset{=}{x}\; (C(x) \Rightarrow (A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x})))$

<2.5> $\forall x\, C(x) \Rightarrow (A(\overset{=}{x}, x) \rightarrow_{Q_\tau} B(\overset{=}{x}, x)) \mapsto\!\leftarrow\!\!\dashv \exists x\, (C(x) \Rightarrow (A(\overset{=}{x}, x) \rightarrow_{Q_\tau} B(\overset{=}{x}, x)))$

<2.6> $\exists x\, C(x) \Rightarrow (A(\overset{=}{x}, x) \rightarrow_{Q_\tau} B(\overset{=}{x}, x)) \mapsto\!\leftarrow\!\!\dashv \forall x\, (C(x) \Rightarrow (A(\overset{=}{x}, x) \rightarrow_{Q_\tau} B(\overset{=}{x}, x)))$

<2.7> $\forall x(C(x) \copyright (A(\overset{=}{x}, x) \rightarrow_{Q_\tau} B(\overset{=}{x}, x))) \mapsto\!\leftarrow\!\!\dashv \forall x\, C(x) \copyright \forall x(A(\overset{=}{x}, x) \rightarrow_{Q_\tau} B(\overset{=}{x}, x)) \copyright \in \{\wedge, \vee\}$

<2.8> $\exists x\, (C(x) \copyright (A(\overset{=}{x}, x) \rightarrow_{Q_\tau} B(\overset{=}{x}, x))) \mapsto\!\leftarrow\!\!\dashv \exists x\, C(x) \copyright \exists x(A(\overset{=}{x}, x) \rightarrow_{Q_\tau} B(\overset{=}{x}, x)) \copyright \in \{\wedge, \vee\}$

The second group of theorems[5] is related to Definition 2 as to reduce the notion of satisfaction in logic to arithmetic testing of a cardinal relation defined by the quantification functions in question. For example, to check if a wff bound by a universal quantifier is true, we check if the size of the model is equal to the size of the universe. As another example, to check if a wff bound by a generalized quantifier Q_{most} is true, we check if the size of the model satisfies the specified cardinal relation (e.g., $\gamma \geq 2/3\ \alpha$). By taking advantage of this concept, we can actually further transform the source assertions, originally in necessary condition (e.g., R_7-R_{12}), into a standardized representation in the form of sufficient condition. The transformation is significant

[4] Proofs of this group of theorems are based on the set-theoretical semantics defined for generalized quantifiers.

[5] Proofs of this group of theorems are based on the symmetric properties of generalized quantifiers (Barwise and Cooper, 1981).

primarily because most subsumption-based reasoning services (Borgida and Brachman, 1993, Bergamaschi, Lodi, and Sartori, 1994, Buchheit, Jeusfeld, Nutt, and Staudt, 1994) that rely on the deduction rule *modus ponens* (McGuinness and Borgida, 1995); that is, $\{\eta \Rightarrow \iota, \eta\} \mapsto \iota$, would require the data source predicate R_i's to be positioned at the right-hand side of $\Rightarrow$, and become the results to be deduced. To discuss the properties of transformation, we need some more understanding of the cardinality behaviors associated with various quantifiers.

<Definition 3>: The **cardinal relation** specified by a quantification function f_τ, is an arithmetic relation of the form $\gamma\ \theta\ \mathrm{Exp}(\gamma, \alpha)$, where θ is an arithmetic operator out of $\{=, >, <, \geq, \leq\}$, and $\mathrm{Exp}(\gamma, \alpha)$ is an arithmetic expression consisting of only two parameters γ and α.
<Definition 4>: A cardinal relation is called **model-independent** if its $\mathrm{Exp}(\gamma, \alpha)$ is a constant.
<Definition 5>: A cardinal relation is called **model-dependent** if its $\mathrm{Exp}(\gamma, \alpha)$ contains at least α.

Table 5 exemplifies the above definitions for those quantifiers that have appeared in this paper. With the notion of model dependence/independence, we can derive Theorem 3.

Quantifier	*Exp(γ, α)*	*Model dependent/independent*	*Category*
$\forall$	$\gamma = \alpha$	dependent	Universal quantifier
$\exists$	$\gamma > 0$	independent	Existential quantifier
Q_{Most}	$\gamma > 2/3\ \alpha$	dependent	Portion quantifier
$Q_{\geq 3}$	$^{\gamma} > 3$	independent	Numeric quantifier
$Q_{=2}$	$\gamma = 2$	independent	Numeric quantifier
$Q_{\text{more-than-1/2}}$	$\gamma > 1/2\ \alpha$	dependent	Fraction quantifier

Table 5 Model dependent and independent quantifiers.

<Theorem 3>: Let ε_1: $A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x})$ be a wff assertion in our language Ω, where A is some data source predicate R. Let Q_τ be a model-independent quantifier, then we can always write ε_1 into ε_2: $B(\overset{=}{x}) \rightarrow_{Q_\tau} A(\overset{=}{x})$ such that $\varepsilon_1 \mapsto\leftarrow\!\dashv \varepsilon_2$.

The above theorem does not apply to model-dependent quantifiers because their cardinal relations involve sizes of the universes, which are reversed after transformation. Now suppose we superscript each $\mathrm{Exp}(\gamma, \alpha)$ to become $\mathrm{Exp}^U(\gamma, \alpha)$ to designate precisely the universe of discourse in question for every particular occurrence of the $\rightarrow_{Q_\tau}$ operator in the source assertions, then the following theorem holds.

<Theorem 4>: Let ε_1: $A(\overset{=}{x}) \rightarrow_{Q_\tau} B(\overset{=}{x})$ be a wff assertion in our language Ω, where A is a data source predicate R. Let Q_τ be a model-dependent quantifier, being only universal, existential, and fraction quantifier, then we can always write ε_1 into ε_2: $B(\overset{=}{x}) \rightarrow_{Q_\tau} \mathrm{Exp}^U(\gamma, \alpha)\ A(\overset{=}{x})$ such that $\varepsilon_1 \mapsto\leftarrow\!\dashv \varepsilon_2$.

Portion quantifier is the only one left out of the above theorems mainly because it requires a further mapping of a qualitative term into a quantitative cardinal relation. In other words, whereas 'Most' can mean 'more than 2/3', it can also be interpreted as 'more than 1/2' or 'more than 3/4', each of which will generate a different cardinal relation. However, once the cardinal relation for a particular portion quantifier is given, then Theorem 2 can be applied. We therefore have the following result.

<Theorem 5>: Let ε_1: $A(\overset{=}{x}) \rightarrow_{Q_\tau} \text{Exp}^{U}(\gamma, \alpha)\ B(\overset{=}{x})$ be a wff assertion in our language Ω, where A is a data source predicate R. Let Q_τ be a portion quantifier and $\text{Exp}^{U}(\gamma, \alpha)$ is given, then we can always write ε_1 into ε_2: $B(\overset{=}{x}) \rightarrow_{Q_\tau} \text{Exp}^{U}(\gamma, \alpha)\ A(\overset{=}{x})$ such that $\varepsilon_1 \rightarrow \leftarrow\!\!\!\dashv \varepsilon_2$.

Consider now a more complex case such that the premise predicate $A(\overset{=}{x}\,;\ x_1 \sim x_i)$ is a conjunctive formula with more than one predicates, one of which is a data source predicate (e.g., R_{12}), then the following theorem is derived.

<Theorem 6>: Let ε_1: $(R(\overset{=}{x}\,;\ x_1 \sim x_i) \wedge C(\overset{=}{x}\,;\ z_1 \sim z_k)) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\ y_1 \sim y_j)$ be a wff assertion in our language Ω, where C is a conjunctive formula, and R is a data source predicate. Let Q_τ be a quantifier, either universal, existential, portion or fraction quantifier, we can always write ε_1 into ε_2: ε_1: $(C(\overset{=}{x}\,;\ z_1 \sim z_k) \wedge B(\overset{=}{x}\,;\ y_1 \sim y_j)) \rightarrow_{Q_\tau} R(\overset{=}{x}\,;\ x_1 \sim x_i)$ such that $\varepsilon_1 \mapsto \leftarrow\!\!\!\dashv \varepsilon_2$.

The third group of theorems[6] is to establish the inferential relations of formulas involving generalized quantifiers and role individual subsumption (i.e., ↑) we discussed in Section 2.1. We only list a partial set of theorems in this group. The value of this group of theorems can be seen from the following inferences stated in English[7].

'If there are more than 5 Taiwan cities in R, then there are more than 5 Asian cities in R'.

'If there are less than 5 Asian cities in R, then there are less then 5 Taiwan cities in R'.

'If all of the Asian cities are in R, then all of the Taiwan cities are in R'.

'If some of the Taiwan cities are in R, then some of the Asian cities are in R'.

<Theorem 7: Deduction with role individual subsumption>: The following properties hold for the language Ω, where θ is the arithmetic operator within the cardinal relation, $\gamma\ \theta\ \text{Exp}(\gamma, \alpha)$,

<7.1> $\{\forall x \uparrow(c, x), A(\overset{=}{x}\,;\ c) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\ c)\} \mapsto A(\overset{=}{x}\,;\ x) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\ x)$ if θ is $\geq$ or $>$

<7.2> $\{\forall x \uparrow(x, c), A(\overset{=}{x}\,;\ c) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\ c)\} \mapsto A(\overset{=}{x}\,;\ x) \rightarrow_{Q_\tau} B(\overset{=}{x}\,;\ x)$ if θ is $\leq$ or $<$

<7.3> $\{\forall x \uparrow(x, c), A(\overset{=}{x}\,;\ c) \rightarrow_{Q_\forall} B(\overset{=}{x}\,;\ c)\} \mapsto A(\overset{=}{x}\,;\ x) \rightarrow_{Q_\forall} B(\overset{=}{x}\,;\ x)$

<7.4> $\{\forall x \uparrow(c, x), A(\overset{=}{x}\,;\ c) \rightarrow_{Q_\exists} B(\overset{=}{x}\,;\ c)\} \mapsto A(\overset{=}{x}\,;\ x) \rightarrow_{Q_\exists} B(\overset{=}{x}\,;\ x)$

The following assertions illustrate the applications of theorems covered in this sub-section to produce assertions of a standardized structure in the form of sufficient condition, which were formulated originally in necessary conditions.

R_7: City (x) $\wedge$ Located(x, Taiwan) $\wedge$ Legal(x, Major)) $\rightarrow_{Q_\forall}$ R_7(x) where $\text{Exp}^{|R_7|}(\gamma, \alpha)$

R_8: City (x) $\wedge$ Located(x, Taiwan) $\wedge$ Legal(x, Major)) $\rightarrow_{Q_{Most}}$ R_8(x) where $\text{Exp}^{|R_8|}(\gamma, \alpha)$

R_9: City (x) $\wedge$ Located(x, Taiwan) $\wedge$ Legal(x, Major)) $\rightarrow_{Q_\exists}$ R_9(x) where $\text{Exp}^{|R_9|}(\gamma, \alpha)$)

R_{10}: City (x) $\wedge$ Located(x, Taiwan) $\wedge$ Legal(x, Major)) $\rightarrow_{Q_{\geq 3}}$ R_{10}(x) where $\text{Exp}^{|R_{10}|}(\gamma, \alpha)$

R_{11}: City (x) $\wedge$ Located(x, Taiwan) $\wedge$ Legal(x, Major)) $\rightarrow_{Q_{More-than-1/2}}$ R_{11}(x) where $\text{Exp}^{|R_{11}|}(\gamma, \alpha)$

R_{12}: (City (x) $\wedge$ Located(x, Asia) $\rightarrow_{Q_{Most}}$ R_{12}(x)) where $\text{Exp}^{|R_{12}|}(\gamma, \alpha)$$\wedge$(City (x) $\wedge$ Located(x, Taiwan)$\rightarrow_{Q_{Most}}$ R_{12}(x) where $\text{Exp}^{|R_{12} \cap \text{City(X)} \cap \text{Located(X, Taiwan)}|}(\gamma, \alpha)$

[6] Proofs of this group of theorems are based on the monotonic properties of generalized quantifiers (Barwise and Cooper, 1981).

[7] The relationship of *role individual subsumption*, ↑, enforces that all Taiwan cities are Asian cities.

4 DISCUSSIONS AND FUTURE WORK

4.1 Summary and application

We have presented a formal representation system for asserting a variety of source contents that prevail on the Internet. Our proposal is motivated by the weakness of description logic in representing sufficient condition for capturing completeness knowledge, and its failure to describe the properties of a concept itself at the set level. We advance a language that exploits the features of generalized quantifiers, and show that this language is expressive in asserting source contents, that otherwise are unable to be stated in other approaches. Our efforts, compared with other related work (Levy, Mendelzon, and Sagiv, 1995, Arens, Chee, Hsu, and Knoblock, 1993), outperforms in allowing data source contents to be characterized in a finer fashion. As a consequence, whereas other systems are eloquent in representing integrity knowledge and therefore provide an effective way to rule out irrelevant sources, our systems are further capable of differentiating those relevant data sources via size differences. In other words, while other competing systems only characterize data sources at the level of relevance, we are able to take a step further to characterize the coverage degree of each relevant source, and represent it using generalized quantifiers and quantification functions.

Our source selection strategy heavily depends on the system's knowledge of the scope and size information of source contents. While such information is sometimes available in web-page format or other electronic forms and public to users' access, those information tends to be given in a very brief fashion; namely, their scope and size are described roughly. We believe that with the current growth in the number of data sources, not before long the pressure from the user community will force source owners to specify data source contents in terms of their scope and size information at a meaningful level of detail. This assumption is especially justified within a single organization with a recognized organizational need to establish a high degree of interoperability among sources, for example, through a data warehouse. Our technology will be able to work with such source specifications and translate them into our proposed language with minimal manual interventions to provide automatic source selection services.

4.2 Source ordering and complexity factors

We are currently using this representation formalism, in particular the size information, to develop a *reasoner* for determining the order of relevant data sources. The order should allow relevant data sources with more possibility answering a request to be given more precedence for selection. In fact, the order among relevant sources can be understood as a generalized view of subsumption where not only the scope information encoded by DL formulas but also the size information embodied by generalized quantifiers is considered as well. In particular, the various generalized quantifiers associated with the satisfied source assertions with respect to a given query should enable us to establish a further prioritization of the relevant sources, thus resulting in a finer-grained concept hierarchy. Indeed, considerations of both scope and size information in a logic-based language involve studying their interactions and demand a new subsumption algorithm consisting of cardinal-number comparisons. Such algorithm should be able to deal with different cases when the sizes of models regarding DL formulas are extensionally given (e.g., model-independent quantifiers) or intensionally related (e.g., model-dependent quantifiers). Theorems covered in Section 3.3 are meant to facilitate that reasoning purpose. More rigorous treatment and detailed algorithm designs as to those issues are currently under our investigation.

The computational behaviors of our proposed language is another important element of our future work. There are two perspectives involved in that effort. One perspective concerns the tractability of determining subsumption of DL formulas incurred in our language. A significant amount of past work in DL has been devoted to this area, and basically come to the conclusion that complexity of DL-based concept languages is essentially determined by the types of constructors that are allowed to appear. Inclusion of certain constructors; for example, *role composition*, will cause subsumption checking to fall beyond the tractability boundary, and lead to intractable computational cost (Brachman and Levesque, 1984), or even undecidability (Schmidt, 1989). Therefore, the main objective of those previous work has been to establish languages that are maximally expressive within the polynomial-time complexity class (Borgida and Patel-Schneider, 1994, Donini, Nutt, Lenzerini, and Nardi, 1991). In that regard, we are interested in being able to show that adding generalized quantifiers conforming to certain properties will still maintain subsumption checking with cardinality comparison in those languages within the tractable complexity class.

The other perspective is related to the computational analysis of cardinality comparison for both extensional and intensional cases. Since the semantics of our source assertions are interpreted as cardinal relations, the source ordering process among relevant sources can be characterized as a constraint solving system where two (or all) data sources are orderable iff the system is arithmetically solvable. In that regard, work on conjunctive query implication (Ullman, 1989), dealing with checking satisfiability of a set of arithmetic equations, offers a sound basis to approach that problem. We are interested in adapting known results in that research area in the context of the first perspective to establish a source description language that is both practically expressive and theoretically tractable.

5 REFERENCES

Anwar, T., Beck, H. and Navathe, S. (1992) Knowledge Mining By Imprecise Querying: A Classification-Based Approach. *IEEE Intl. Conf. On Data Eng.*, 622-630.

Arens, Y., Chee, C., Hsu, C. and Knoblock, C. (1993) Retrieving and Integrating Data from Multiple Information Sources. *Intl. Journal of Intelligent and Cooperative Information Systems*, 2(2), 127-158.

Baader, F. (1990) Terminological Cycles in KL-ONE-based Knowledge Representation Languages. *AAAI-90*, 621-626.

Baader, F. (1996) A Formal Definition for the Expressive Power of Terminological Knowledge Representation Languages. *Journal of Logic Computation,* 6(1), 33-54.

Borgida, A. and Brachman, R. (1993) Loading Data Into Description Reasoner. *SIGMOD Intl. Conf. on Mgnt. of Data*, 217-226.

Beneventano, D., Bergamanschi, S. and Lodi, C. (1994) Terminological Logics for Schema Design and Querying Processing in OODBs. 1st *KRDB-94*.

Borgida, A., Brachman, J. and McGuinness, D. (1989) CLASSIC: A Structural Data Model For Objects. *SIGMOD Intl. Conf. on Mgnt. of Data*, 58-67.

Barwise J. and Cooper, R. (1981) Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, 4, 159-219.

Benthem, J. (1982) Questions About Quantifiers. *The Journal of Symbolic Logic*, 4(2), 443-466.

Brachman, R., Fikes, R. and Levesque, H. (1983) Krypton: A Functional Approach to Knowledge Representation. *IEEE Computer*, 16(10), 67-73.

Beck, H., Gala, S. and Navathe, S. (1989) Classification As a Query Processing Technique in the CANDIDE Semantic Data Model. *IEEE Intl. Conf. on Data Eng.*, 572-581.

Buchheit, M., Jeusfeld, M., Nutt, W. and Staudt, M. (1994) Subsumption Between Queries to Object-oriented Databases. *Information Systems*, 33-54.

Bergamaschi, S. and Sartori, C. (1992) On Taxonomic Reasoning in Conceptual Design. *ACM TODS*, 17(3), 385-442.

Bergamaschi, S., Lodi, S. and Sartori, C. (1994) The E/S Knowledge Representation System. *Data & Knowledge Engineering*, 14, 81-115.

Brachman, R. and Levesque, H. (1984) The Tractability of Subsumption in Frame-based Description Languages. *AAAI'84*, 34-37.

Borgida, A. (1995) Description Logics in Data Management. *IEEE TKDE*, 7(5), 671-682.

Bresciani, P. (1995) Querying Databases from Description Logics. 2nd *KRDB-95*.

Bresciani, P. (1996) Some Research Trend in KR & DB. 3 rd *KRDB-96*.

Borgida, A. and Patel-Schneider, P. (1994) A Semantics and Complete Algorithm for Subsumption in the CLASSIC Description Logic. *Journal of Artificial Intelligence Research*, Vol. 1, 277-308.

Brachman, R. and Schmolze, J. (1985) An Overview of the KL-One Knowledge Representation System. *Cognitive Science*, 9(2), 171-216.

Chen, P. (1976) The Entity-Relationship Model: Toward A Unified View of Data. *ACM TODS*, 1(1), 9-36.

Calvanese, D., Lenzerini, M. and Nardi, D. (1992) A Unified Framework for Class-Based Representation Formalisms. *KR-92*, 109-120.

Donini, F., Nutt, W., Lenzerini, M. and Nardi, D. (1991) Tractable Concept Languages. *IJCAI-91*, 458-468.

Date, C. J. (1986) *An Introduction to Database Systems*. Addison Wesley Publishing Company.

Devanbu, P. (1993) Translating Description Logics into Information Server Queries. *2nd Intl. Conf. on Information and Knowledge Management*.

Doyle, J. and Patil, R. (1991) Two Theses of Knowledge Representation: Language Restrictions, Taxonomic Classification, and The Utility of Representation Services. *Artificial Intelligence*, 48, 261-297.

Enderson, H. (1971) *A Mathematical Introduction to Logic*. Academic Press Inc.

Etzioni, O., Gloden, K. and Weld, D. (1994) Tractable Closed World Reasoning With Updates. *KR-94*, 178-189.

Etzioni, O., Gloden, K. and Weld, D. (1994) A Softbot-based Interface to The Internet. *CACM*, 37(7), 72-76.

Frege, G. (1968) The Foundation of Arithmetic. Tran. J. L. Evaston I11: Northwestern U. Press.

Giacomo, D. and Lenzerini, G. (1996) Tbox and Abox Reasoning in Expressive Description Logics. *KR-96*, 316-327.

Hollunder, B. and Baader, F. (1994) Qualifying Number Restriction in Concept Languages. *KR-94*, 335-346.

Hull, R. and King, R. (1987) Semantic Database Modeling: Survey, Applications, and Research Issues. *ACM Computing Survey*, 19(4), 201-260.

Hsu, P. and Parker, D. (1995) Improving SQL with Generalized Quantifiers. *IEEE Intl. Conf. on Data Eng.*, 298-305.

Kent, W. (1978) Data and Reality. North-Holland Publishing Company.

Kessel, T., Rousselot, F., Schlick, M. and Stern, O. (1995) Use of DL within the Framework of DBMS. 2nd *KRDB-95*.

Knoblock, C., Yigal, A. and Hsu, C. (1994) Cooperating Agents for Information Retrieval. *2nd Intl. Conf. on Cooperative Information Systems.*

Levy, A., Mendelzon, Y. and Sagiv, A. (1995) Answering Queries Using Views. *PODS-95*, 95-105.

Levy, A., Rajarman, A. and Ordille, J. (1996) Query-Answering Algorithms for Information Agents. *AAAI-96*, 40-47.

Levy, A., Rajarman, A. and Ordille, J. (1996) Querying Heterogeneous Information Sources Using Source Descriptions. *VLDB-96.*

Levy, A., Srivastava, D. and Kirt, T. (1995) Data Model and Query Evaluation in Global Information systems. *Journal of Intelligent Information Systems*, 5(2), 121-143.

MacGregor, R. (1994) A Description Classifier for The Predicate Calculus. *AAAI-94*, 213-220.

MacGregor, R. and Bates, R. (1987) The LOOM Knowledge Representation Language. Technical Report ISI/RS-87-188, USC/ISI.

Mays, W., Tyler, S., McCuire, J. and Schlossberg, J. (1987) Organizing Knowledge in a Complex Financial Domain. *IEEE Expert*, 2(3), 61-70.

McGuinness, D. and Borgida, A. (1995) Explaining Subsumption in Description Logics. *IJCAI-95*, 816-821.

Mostowski, A. (1957) On a Generalization of Quantifiers. *Fundamenta Mathematicae*, 44, 12-36.

Motro, A. (1986) Completeness Information and Its Applications to Query Processing. *VLDB-86*, 170-178.

Motro, A. (1989) Integrity = Validity + Completeness. *ACM TODS*, 14(4), 480-502.

Patel-Schnider, P., Brachman, R. and Levesque, H. (1984) ARGON: Knowledge Representation Meets Information Retrieval. *Intl. Conf. on AI Application*, 280-286.

Patel-Schneider, P. and Swartout, B. (1993) Description Logic Specification from the KRSS Effort.

Quantz, J., Schmitz, B. and Kussner, B. (1994) Using Description Logics for Disambiguation in Natural Language Processing. *DL-94.*

Quantz, J. (1992) How to Fit Generalized Quantifiers Into Terminological Logics. *ECAI-92*, 543-547.

Schmidt, M. (1989) Subsumption in KL-ONE is Undecidable. *KR-89*, 421-431.

Sher, G. (1991) The Bounds of Logic: A Generalized Viewpoint. The MIT Press.

Westerstahl, D. (1989) Quantifiers in Formal and Natural Languages. *Handbook of Philosophical Logic*, 1-131.

Ullman, J. (1989) *Principles of Database and Knowledge-Base Systems.* Vol. 1, Computer Science Press, New York.

6 BIOGRAPHY

S. Y. Tu: Steven Yi-cheng Tu is currently a doctoral student majoring in Information Technology at the Sloan School of Management of Massachusetts Institute of Technology. He holds a Bachelor degree in Business Administration, and a M.S. degree in Computer Science. His research interests include database and knowledge-based systems, and most recently logic-based and object-based database systems. He has published papers in the local computer science journals, and in the proceedings of international and national MIS conferences. He plans to pursue an academic career at Taiwan upon completion of his graduate study.

S. Madnick: Stuart Madnick is the John Norris Maguire Professor of Information Technology and Leaders for Manufacturing Professor of Management Science at the MIT Sloan School of

Management. He is also an affiliate member of the MIT Laboratory for Computer Science and a member of the Executive Committee of the MIT Center for Information Systems Research. His current research interests include connectivity among disparate distributed information systems, database technology, and software project management. He is the author or co-author of over 200 books, articles, or reports on these subjects, including the classic textbook, Operating Systems (McGraw-Hill), and the book, The Dynamics of Software Development (Prentice-Hall). He has been active in industry, making significant contributions as one of the key designers and developers of projects such as IBM's VM/370 operating system and Lockheed's DIALOG information retrieval system.