Distributed Description Logics:

– preliminary investigations –

Alex Borgida
Dept. of Computer Science
Rutgers University, NJ 08855, USA
email: borgida@cs.rutgers.edu

Luciano Serafini
ITC-IRST
38050 Povo, Trento, Italy
email: serafini@itc.it

Abstract

Information integration has been, and remains one of the major challenges of information processing, and is well supported by description logics. We illustrate the need for a more refined approach in those cases where the original sources form a loosely federated information system, with each one wishing to maintain its own independent view of the world. We formalize this using the notion of Distributed Description Logics, where there are *directional* mappings between the local sources, including mappings between individuals.

1 Introduction

A significant problem of modern information management is the integration of information from multiple sources. The standard version presumes a framework where users are accessing through a single interface data from several information sources (local ISs), which can include databases, web data, files, etc. The important goal here is making the users unaware of the original source of the information, thus making the interface uniform. This is usually achieved through a global (conceptual) schema that is queried by users. Local ISs are then related to this by a variety of techniques ("local as view", "global as view"), and query answering consists of identifying relevant ISs, translating the user's query into collections of queries over local ISs, and collating the answers.

A somewhat different, but related, approach is one which preserves the identity of each local IS and its user interface. However, the local system wishes to import information available in other sources, which are related to it directly through bilateral assertions. This approach is more appropriate for so-called "federated/distributed" information system, where the local ISs maintain a degree of autonomy, and it is this framework that we will be exploring.

No matter how inter-dependencies between the IS are expressed, there are some problems whose solution is of general interest, including *verifying the consistency of the integration*, and *query processing*. As argued in [2], both these tasks can be supported by representing the ISs and their integration in some logic, and then reasoning about the result.

We start by assuming that we have a collection of information systems IS_i, each described by a Decription Logic (DL) theory T_i, ("T-box"). To express relationships between the IS_i, the pioneering work of Catarci and Lenzerini [2] proposed the continued use of description logics. For example, $GradStudent_2 \sqsubseteq_{int} Student_1$ was meant to indicate that every graduate student in the part of the world described by IS_2 was also a student in the overlapping part of the world described by IS_1.

However, the semantics in [2] indicates that interschema assertions only have an effect on those *individuals that are shared* between the respective IS domains (i.e., a uniform ontology of individuals is assumed). The following examples illustrate the need for a more refined approach.

Example 1.1. *Suppose IS_1 has information about couples/families, while IS_2 has information about persons. There are clear relationships between the information in the two IS, e.g., if a couple has an associated address, then the husband and wife who make up the couple can be deduced to have the same addresses.*

The problem here is that IS_1 contains information about individuals that are *abstractions* over individuals in IS_2. Similar examples occur in any situation where the so-called "materialization abstraction" [5] occurs (e.g., a play vs. a production of the play vs. a particular performance). In general, we need to express *a relationship between the individuals* in the two domains, e.g., between couple23 in IS_1, and each of Gianni and Mary, say, in IS_2. This relationship must in fact be *directed*, as shown by the case of correspondences between Lire and Dollars, which are not converses due to the commission charged by banks.

Example 1.2. *Suppose C1, D and C2 are 3 increasingly difficult courses on related topics. University u1 offers C1 and C2, while university u2 offers D. The universities allow a course x to be substituted for another y, even as a transfer,* if *x is harder than y,* and *covers most of the material of y (say 80%). U1 may decide to treat D as equivalent to C1; on the other hand, according to u2, D is only equivalent to C2, since C1 is weaker than D. If courses are viewed as collections of students having taken them, this can be rephrased in terms of class subsumption: according to u1, class C1 subsumes D, while according to u2, class D subsumes C2; but u1 might not view C2 as a subclass of C1, since these might disagree on more than 80% of the material.*

The above example shows that even more general relationships between ISs, such as subsumption, should be directed. (Another reason for this may be that IS_1 trusts IS_2, and is willing to import its data, but not conversely.)

To handle these issues we propose a generalized framework, called *distributed description logics* (DDL), inspired by Distributed First Order Logic [3]. A distributed T-box consists of local T-boxes T_i, described using ordinary DLs, and *bridge-rules* relating them. We limit bridge rules here so that they only relate ISs pairwise. In order to support directionality, the bridge rules in a set $\mathfrak{B}_{jk}$ will be viewed as describing "flow of information" from IS_j to IS_k *from the point of view of* IS_k (i.e., IS_k "importing" information from IS_j), and hence $\mathfrak{B}_{jk}$ may be different from $\mathfrak{B}_{kj}$.

Since our primary aim is to resolve issues dealing with the correspondence of individuals between two IS, bridge rules will be concerned with this aspect. Based on studies in [3], here are some types of constraints on correspondence relationships that one might like to express using bridge rules:

1. Every A-object in T_1 corresponds only to G-objects in T_2.
2. All H-objects in T_2 have a corresponding A-object in T_1.
3. Each A-object has at least/at most one corresponding object in T_2.
4. The correspondence relation from T_1 to T_2 is the identity relationship.

In the remainder of the paper we first present the formal definition and model theory of DDL. We then discuss some of their properties, including a mapping from a DDL to a "global" DL, which, under certain circumstances, allows deduction in DDL to be simulated in ordinary DL. We also investigate the desired behaviour of DDL in the presence of inconsistency.

2 Formal Definitions

We assume the reader is familiar with the definitions of the syntax and semantics of description logics, using interpretations $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ that assign subsets of $\Delta^{\mathcal{I}}$ to atomic concepts, and subsets of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ to atomic roles.

To begin our formal definitions let I be a nonempty set of indexes, and $\mathcal{DL}_i$ be description logics for every $i \in I$. Furthermore, let T_i be T-boxes in $\mathcal{DL}_i$.

To avoid confusion, we want the set of descriptions in each $\mathcal{DL}_i$ to be distinct. To do so, we can label each description E in $\mathcal{DL}_i$ with its index i (written as $i : E$). However, when talking about subsumption within a single IS_i, we will use the more readable $i\!:\!A \sqsubseteq B$, instead of the more formal $i\!:\!A \sqsubseteq i\!:\!B$.

Definition 2.1. Given concepts C and G of $\mathcal{DL}_i$ and $\mathcal{DL}_j$ respectively, a *bridge* *rule* from i to j is an expression of the following two forms:

$$i:C \xrightarrow{\sqsubseteq} j:G \qquad \text{called an } \textit{into rule}$$
$$i:C \xrightarrow{\sqsupseteq} j:G \qquad \text{called an } \textit{onto rule}$$

A *distributed description logic* $\mathcal{D}$ consists of a set of description logics $\{\mathcal{DL}_i\}_{i\in I}$, which is intended to constrain, in the manner made precise below, the various information sources and their connections.

A *distributed T-box* (DTB) $\mathfrak{T} = \langle\{T_i\}_{i\in I}, \mathfrak{B}\rangle$ for a DDL $\{\mathcal{DL}_i\}_{i\in I}$, consists of a set of T-boxes $\{T_i\}_{i\in I}$, and a set $\mathfrak{B} = \{\mathfrak{B}_{ij}\}$ of bridge rules from i to j for every $i \neq j \in I$. For every $k \in I$, all descriptions in T_k must be in the corresponding language $\mathcal{DL}_k$, and for every bridge rule $i : A \xrightarrow{\sqsubseteq} j : B$ or $i : A \xrightarrow{\sqsupseteq} j : B$ in $\mathfrak{B}_{ij}$, the concepts A and B must be in the languages $\mathcal{DL}_i$ and $\mathcal{DL}_j$ respectively.

The set $\{\mathcal{DL}_i\}_{i\in I}$ is in a sense the "type signature" of the DTB, and in our examples it will be clear from the context. It is however important to realize that distributed description logics provide an interesting opportunity to study hybrid/multi-language reasoning, where different local IS use *different DL's* to express their information, thus leading to possibly different computational properties for the combined DDL.

In Example 1.1, one would have the bridge rule $1{:}\texttt{COUPLE} \xrightarrow{\sqsubseteq} 2{:}\texttt{PERSON}$ to indicate that every couple has corresponding persons, but would not normally include the bridge rule $1{:}\texttt{COUPLE} \xrightarrow{\sqsupseteq} 2{:}\texttt{PERSON}$ because there may be unmarried persons who are not part of any couple. The specific correspondence of Gianni and Mary to couple23 can be expressed by bridge rules $1 : \{\mathsf{couple23}\} \xrightarrow{\sqsubseteq} 2 : \{\mathsf{Gianni, Mary}\}$ and $1{:}\{\mathsf{couple23}\} \xrightarrow{\sqsupseteq} 2{:}\{\mathsf{Gianni, Mary}\}$, if the description logics support concepts formed by enumeration. Otherwise, we need to resort to new kinds of assertions about individual mappings – see Section 6,

The semantics of DDL is based on the notion of distributed interpretation, which uses relations r_{ij} to connect the domains of the component IS's:

Definition 2.2. A *distributed interpretation* $\mathfrak{I} = \langle\{\mathcal{I}_i\}_{i\in I}, r\rangle$ of $\mathfrak{T}$ consists of interpretations $\mathcal{I}_i$ for $\mathcal{DL}_i$ over domain $\Delta^{\mathcal{I}_i}$, and a function r associating to each $i, j \in I$ a binary relation $r_{ij} \subseteq \Delta^{\mathcal{I}_i} \times \Delta^{\mathcal{I}_j}$. We use $r_{ij}(d)$ to denote $\{d' \in \Delta^{\mathcal{I}_j} \mid \langle d, d'\rangle \in r_{ij}\}$, and for any $D \subseteq \Delta^{\mathcal{I}_i}$, we use $r_{ij}(D)$ to denote $\bigcup_{d\in D} r_{ij}(d)$.

Definition 2.3. A distributed interpretation $\mathfrak{I}$ *d-satisfies* (written $\mathfrak{I} \models_d$) the elements of a DTB $\mathfrak{T} = \langle\{T_i\}_{i\in I}, \{\mathfrak{B}_{ij}\}\rangle$ according to the following clauses: For every $i, j \in I$

1. $\mathfrak{I} \models_d i{:}B \xrightarrow{\sqsubseteq} j{:}G$, if $r_{ij}(B^{\mathcal{I}_i}) \subseteq G^{\mathcal{I}_j}$;
2. $\mathfrak{I} \models_d i{:}A \xrightarrow{\sqsupseteq} j{:}H$, if $r_{ij}(A^{\mathcal{I}_i}) \supseteq H^{\mathcal{I}_j}$;

3. $\mathfrak{I} \models_d i : A \sqsubseteq B$, if $\mathcal{I}_i \models i : A \sqsubseteq B$,
4. $\mathfrak{I} \models_d T_i$ iff $\mathfrak{I} \models_d i : A \sqsubseteq B$ for all $A \sqsubseteq B$ in T_i.
5. $\mathfrak{I} \models_d \mathfrak{T}$ for $\mathfrak{T} = \langle \{T_i\}_{i \in I}, \mathfrak{B} \rangle$ if, for every $i \in I$, $\mathfrak{I} \models_d T_i$, and $\mathfrak{I}$ d-satisfies every bridge rule in $\mathfrak{B}$.

Finally, $\mathfrak{T} \models_d i : C \sqsubseteq D$ if, for every distributed interpretation $\mathfrak{I}$, $\mathfrak{I} \models_d \mathfrak{T}$ implies $\mathfrak{I} \models_d i : C \sqsubseteq D$.

Example 2.1. *Consider two T-boxes T_l and T_s. T_l describes the schema of the IRST library database. It contains a concept* `BOOK` *for all the books belonging to the library. Note that the library can contain multiple copies of the same book, which are then distinct invidividuals in* `BOOK`. *T_l contains also the concept* `PERSON` *for people who can borrow books from the library, and the role* `taken_by` *that records who is currently in possession of a book. Finally, T_l contains the defined concept* `BOOK_ON_SHELF`, *representing all the books that are not on currently on loan at this library.* `BOOK_ON_SHELF` *is defined as follows:*

$$\texttt{BOOK\_ON\_SHELF} \equiv \texttt{BOOK} \sqcap \neg \exists \texttt{taken\_by.PERSON}$$

The T-box T_s describes the simple schema of a personal information base used by a student to keep track of the books available in the libraries of Trento. T_s includes the concept `BOOK`, *which in this case does not distinguish between different copies of the same book, since all that is relevant is whether one can get* some *copy of the book by going to the appropriate library. T_s also refers to* `LIBRARY` *– the set of accessible libraries, plus the role* `located_at`, *which associates with a book the libraries where it is available. Clearly, these information bases are related. Indeed, the information about (available) books in the student personal database is supposed to be imported from the libraries' database. This relation is directional, as (hopefully!) no information leaves the student personal database. The bridge rules that describe the relation between T_l and T_s are the following*

$$l : \texttt{BOOK} \xrightarrow{\sqsubseteq} s : \texttt{BOOK} \tag{1}$$

$$l : \texttt{BOOK\_ON\_SHELF} \xrightarrow{\sqsupseteq} s : \exists \texttt{located\_at.}\{\texttt{"Irst"}\} \tag{2}$$

Bridge rule (1) formalizes the fact that some books in the IRST library are copies of books of potential interest to the student (i.e., instances of s : `BOOK`*). This should not be confused with the requirement that* all *books of the IRST library be mapped to some book in the student personal base, which is not given here, and would hold only if the domain relation r_{ls} was defined on all the elements of the domain of T_l which are* `BOOK`*s. Bridge rule (2) formalizes the fact that the student knows that something is located at the IRST library only if it is a book that is not on loan there. (This means that the only way that a student can learn about availability of material at the IRST library is by accessing it.)*

Let $\mathfrak{T}_{sl} = \langle T_l, T_s, \mathfrak{B}_{sl} = \{(1)(2)\}\rangle$ be the distributed T-box that formalizes the student and IRST libraries database with their relation. An example of distributed interpretation for $\mathfrak{T}_{sl}$ is $\mathfrak{I}_{sl}$, as described in Figure 1.

$$
\begin{array}{rcl}
\Delta^{\mathcal{I}_l} & = & \{\texttt{Complexity\_Theory(1)}, \texttt{Complexity\_Theory(2)}, \texttt{DB\_Pples}, \texttt{Mario}\} \\
\texttt{BOOK}^{\mathcal{I}_l} & = & \{\texttt{Complexity\_Theory(1)}, \texttt{Complexity\_Theory(2)}, \texttt{DB\_Pples}\} \\
\texttt{PERSON}^{\mathcal{I}_l} & = & \{\texttt{Mario}\} \\
\texttt{taken\_by}^{\mathcal{I}_l} & = & \{\langle \texttt{Complexity\_Theory(1)}, \texttt{Mario}\rangle\} \\
\Delta^{\mathcal{I}_s} & = & \{\texttt{Complexity\_Theory}, \texttt{The\_society\_of\_mind}, \texttt{"Irst"}, \texttt{"Univ TN"}\} \\
\texttt{BOOK}^{\mathcal{I}_s} & = & \{\texttt{Complexity\_Theory}, \texttt{The\_society\_of\_mind}\} \\
\texttt{located\_at}^{\mathcal{I}_s} & = & \left\{ \begin{array}{c} \langle \texttt{Complexity\_Theory}, \texttt{"Irst"}\rangle \\ \langle \texttt{The\_society\_of\_mind}, \texttt{"Univ TN"}\rangle \end{array} \right\} \\
r_{ls} & = & \left\{ \begin{array}{c} \langle \texttt{Complexity\_Theory(1)}, \texttt{Complexity\_Theory}\rangle \\ \langle \texttt{Complexity\_Theory(2)}, \texttt{Complexity\_Theory}\rangle \end{array} \right\}
\end{array}
$$

Figure 1: Example of distributed interpretation for $\mathfrak{T}_{sl}$

$\mathfrak{I}_{sl}$ satisfies bridge rule (1); indeed, $r_{ls}(\texttt{BOOK}^{\mathcal{I}_l}) = \{\texttt{Complexity\_Theory}\} \subseteq \texttt{BOOK}^{\mathcal{I}_s} = \{\texttt{Complexity\_Theory}, \texttt{The\_society\_of\_mind}\}$; $\mathfrak{I}_{sl}$ also satisfies bridge rule (2); indeed, $r_{ls}(\texttt{BOOK\_ON\_SHELF}^{\mathcal{I}_l}) = r_{ls}(\{\texttt{Complexity\_Theory(2)}\}) = \{\texttt{Complexity\_Theory}\} \supseteq (\exists \texttt{located\_at}.\{\texttt{"Irst"}\})^{\mathcal{I}_s} = \{\texttt{Complexity\_Theory}\}$. Note also that the bridge rule $l : \texttt{BOOK} \xrightarrow{\sqsubseteq} s : \texttt{BOOK}$ is satisfied even if $\texttt{BOOK}^{\mathcal{I}_l}$ is not contained in $\texttt{BOOK}^{\mathcal{I}_s}$. Finally, $s : \exists \texttt{located\_at}.\{\texttt{"Irst"}\} \sqsubseteq \texttt{BOOK}$ is a logical consequence of $\mathfrak{T}_{sl}$. I.e. the bridge rules allow us to infer the fact in T_s that anything that is located at the IRST library must be a book, and hence an instance of a concept $\texttt{AV\_BOOK}$, which the student might have defined in T_s as

$$\texttt{AV\_BOOK} \equiv \texttt{BOOK} \sqcap \exists \texttt{located\_at}.\texttt{LIBRARY}$$

for the purposes of finding quickly books that are available to her.

3 Some properties of DDL

We will consider the simplest kinds of DTB $\mathfrak{T}_{12}$, of the form $\langle T_1, T_2, \mathfrak{B}_{12}\rangle$, with only two T-boxes T_1 and T_2, and a single set of bridge rules $\mathfrak{B}_{12}$ from 1 to 2.

We present a series of statements describing some of the properties of DDL. (The proofs are quite straightforward.)

First, all local information is available for reasoning in the distributed system. Indeed, if $T_i \models i{:}X \sqsubseteq Y$, then $\mathfrak{T}_{12} \models_d i{:}X \sqsubseteq Y$.

Next, if the set of bridge rules is empty, then no information can pass between IS_1 and IS_2 (unless $\mathfrak{T}_{12}$ is inconsistent): if $\langle T_1, T_2, \emptyset\rangle \models_d i : X \sqsubseteq Y$ and T_j is consistent, then $T_i \models i{:}X \sqsubseteq Y$, for $i, j \in \{1, 2\}$.

The following displays a pattern of how subsumption information can be passed from IS_1 to IS_2: if $T_1 \models 1:A \sqsubseteq B$, and $\mathfrak{B}_{12}$ contains the bridge rules $1\!:\!B \xrightarrow{\sqsubseteq} 2\!:\!H$, and $1\!:\!A \xrightarrow{\sqsupseteq} 2\!:\!G$, then $\mathfrak{T}_{12} \models_d 2\!:\!G \sqsubseteq H$. (The intuition behind this is the containment $G^{\mathcal{I}_2} \subseteq r_{ij}(A^{\mathcal{I}_1}) \subseteq r_{ij}(B^{\mathcal{I}_1}) \subseteq H^{\mathcal{I}_2}$.)

Finally, if $\mathfrak{B}_{12}$ contains only into bridge rules, then $\mathfrak{T}_{12} \models_d i:X \sqsubseteq Y$ iff $\langle T_1, T_2, \emptyset\rangle \models_d i:X \sqsubseteq Y$. A similar, though not identical, results holds when there are only onto rules.

One of the desired characteristics of DDL is the *unidirectionality* of information flow: information in IS_2, and the bridge rules from IS_1 to IS_2 set up to augment it, should not affect, by themselves, reasoning in IS_1. Unfortunately, as in most logics, if IS_2 is inconsistent, then anything can be derived, even concerning IS_1. (We shall consider this problem in Section 5). In some DL, "back-flow" can occur even with consistent T-boxes; e.g., if $\top_{role}$ denotes the "universal role", which has interpretation $\Delta^{\mathcal{I}_1} \times \Delta^{\mathcal{I}_1}$ for every $\mathcal{I}_1$, then $\left\langle \emptyset, \emptyset, \{1\!:\!A \xrightarrow{\sqsupseteq} 2\!:\!\top\}\right\rangle \models_d 1:\top \sqsubseteq \exists\top_{role}.A$ because, in order to satisfy $1\!:\!A \xrightarrow{\sqsupseteq} 2\!:\!\top$, any distributed interpretation $\langle \mathcal{I}_1, \mathcal{I}_2, r_{12}\rangle$ must have $r_{12}(A^{\mathcal{I}_1}) \neq \emptyset$. For restricted cases, however, we get the desired result:

Proposition 3.1 (no "back-flow" for $\mathcal{SHIQ}$[4]). *Let $\mathfrak{T}_{12}$ be a consistent DTB $\langle T_1, T_2, \mathfrak{B}_{12}\rangle$, where T_1 and T_2 contain only subsumptions from the DL $\mathcal{SHIQ}$. Then $\mathfrak{T}_{12} \models_d 1\!:\!A \sqsubseteq B$ iff $T_1 \models A \sqsubseteq B$.*

4 Relating DDL and ordinary DL

The following provides in certain circumstances a connection between DDL and ordinary DLs, supplying both proof techniques and algorithms for reasoning in some types of DDL.

Given a family of description logics $\{\mathcal{DL}_i\}_{i\in I}$, let the global description logic $\mathcal{GDL}$ have all the primitive concepts and roles of each $\mathcal{DL}_i$. For any primitive concept [role] A [R] of $\mathcal{DL}_i$, let $i:A$ [$i:R$] be a primitive concept [role] of $\mathcal{GDL}$. This language permits at least all composite descriptions of $\mathcal{DL}_i$. $\mathcal{GDL}$ has special top and bottom concepts, $\top_g$ and $\bot_g$, and special role symbols R_{ij}, which will be used to simulate the domain relations.
First, define a mapping $\#$ from concepts/roles of $\mathcal{DL}_i$ to $\mathcal{GDL}$ as follows:

1. $\#(i\!:\!M) = i\!:\!M$ for primitive concepts/roles M;
2. if ρ is a concept constructor taking k arguments, then $\#(\rho(M_1,\ldots,M_k)) = i\!:\!\top \sqcap \rho(\#(M_1),\ldots,\#(M_k))$.
3. analogously for role constructors, but substituting $i\!:\!\top_{role}$ for $i\!:\!\top$.

Definition 4.1. Applying $\#()$ to a DTB $\mathfrak{T} = \langle\{T_i\}_{i\in I}, \mathfrak{B}\rangle$, yields a T-box $\#(\mathfrak{T})$ in the language $\mathcal{GDL}$, consisting of the following axioms:

1. $\#(i\!:\!A) \sqsubseteq \#(i\!:\!B)$ for all $i\!:\!A \sqsubseteq B \;\in T_i$; *(copies of local axioms)*
2. $\#(i\!:\!A) \sqsubseteq \forall \mathsf{R}_{ij}.\#(j\!:\!G)$ for every into bridge rule $i\!:\!A \xrightarrow{\sqsubseteq} j\!:\!G \in \mathfrak{B}$;
3. $\#(j\!:\!H) \sqsubseteq \exists \mathsf{R}_{ij}^{-}.\#(i\!:\!A)$ for every onto bridge rule $i\!:\!A \xrightarrow{\sqsupseteq} j\!:\!G \in \mathfrak{B}$;
4. $\top_g \sqsubseteq \forall \mathsf{R}_{ij}.j:\top$ *(the range of R_{ij} is $\Delta^{\mathcal{I}_j}$)* and $\neg(i:\top) \sqsubseteq \forall \mathsf{R}_{ij}.\bot_g$ *(R_{ij} is undefined outside $\Delta^{\mathcal{I}_i}$)*
5. $i\!:\!\bot \sqsubseteq \bot_g$; *(all the bottom symbols denote the empty set)*
6. $i\!:\!A \sqsubseteq i\!:\!\top$, for every atomic concept A of $\mathcal{DL}_i$;
7. $i\!:\!\top \sqsubseteq \forall i\!:\!R.i\!:\!\top$ for every role R of $\mathcal{DL}_i$ *(the range of $i\!:\!R$ is in $\Delta^{\mathcal{I}_i}$)*
8. $\neg(i\!:\!\top) \sqsubseteq \forall i\!:\!R.\bot_g$ *($i\!:\!R$ is undefined outside $\Delta^{\mathcal{I}_i}$)*

(The last three axioms make sure that "$i\!:\!\top$" behaves properly.)

Theorem 4.1. *$\#(\mathfrak{T}) \models \#(i\!:\!X) \sqsubseteq \#(i\!:\!Y)$ if and only if $\mathfrak{T} \models_d i\!:\!X \sqsubseteq Y$.*

Using Theorem 4.1, we can obtain reasoners for a variety of DDL.

Proposition 4.2. *A DDL such that all $\mathcal{DL}_i$ are contained in some decidable description logic $\mathcal{DL}_0$, which supports (i) qualified existential restriction, and (ii) arbitrary subsumption assertions in T-boxes, can use the decision procedure of $\mathcal{DL}_0$ to decide unsatisfiability and d-entailment.*

Proof. We know that reasoning in $\mathfrak{T}_{12}$ is equivalent to ordinary DL reasoning in $\#(\mathfrak{T}_{12})$. The proof relies on two observations: (a) (by design) every axiom in $\#(\mathfrak{T}_{12})$ involving R_{ij} is either of form $\alpha \sqsubseteq \forall \mathsf{R}_{ij}.\beta$ or $\exists \mathsf{R}_{ij}^{-}.\delta \sqsubseteq \gamma$; (b) an axiom of the form $\alpha \sqsubseteq \forall p.\beta$ is equivalent to $\exists p^{-}.\alpha \sqsubseteq \beta$. This allows all axioms involving R_{ij} to be rewritten to involve only qualified existentials over R_{ij}^{-}, at which point we can replace R_{ij}^{-} by some new role Q_{ij}. This removes the need for inverse roles and universal restrictions. □

We get as corollaries that DDLs with $\mathcal{DL}_i$ that are in $\mathcal{ALCNR}$ or $\mathcal{SHIQ}$ [4] can use their reasoners for determining $\models_d$.

The absence of *nested* existential restricitions on R_{ij} raises the possibility that DDL with simple $\mathcal{DL}_i$ can perform reasoning even more quickly. The simplest DL is one that has only atomic concepts. We have the following:

Proposition 4.3. *Given a DTB, $\mathfrak{T}_{12} = \langle T_1, T_2, B_{12}\rangle$, where T_1, T_2, and B_{12} involve only atomic concepts. Then $\mathfrak{T} \models_d 2\!:\!G \sqsubseteq H$ if and only if $\#(\mathfrak{T}) \vdash 2\!:\!G \sqsubseteq 2\!:\!H$, where the inference rules are*

$$\bot \sqsubseteq \alpha \qquad \alpha \sqsubseteq \top \qquad \exists p.\bot \sqsubseteq \bot \qquad X \sqsubseteq X \qquad \frac{X \sqsubseteq Y\,,\; Y \sqsubseteq Z}{X \sqsubseteq Z} \qquad \frac{X \sqsubseteq Y}{\exists p.X \sqsubseteq \exists p.Y}$$

(Our proof of this completeness result relies on an analysis of the rewrite rules for $\mathcal{SHIQ}$ [4] as they apply in this case.)

This indicates that the cost of reasoning in such DDL is the same as that of reasoning with the corresponding DL, namely the cost of computing transitive closure.

Note that above inference rules are insufficient for more complex languages. For example, $1:(C \sqcup D) \xrightarrow{\sqsupseteq} 2:H$, $1:C \xrightarrow{\sqsubseteq} 2:G$, and $1:D \xrightarrow{\sqsubseteq} 2:G$, should allow us to conclude $2:H \sqsubseteq G$.

5 Unsatisfiability in DDL

From the definition, it can be seen that a DTB $\mathfrak{T}$ is unsatisfiable if each distributed interpretation $\mathfrak{I}$ does not satisfy either some local T-boxes, some bridge rules (for instance $1:\bot \xrightarrow{\sqsupseteq} 2:\top$), or some combinations thereof (for instance $\langle\{A \sqsubseteq \bot\}, \{\top \sqsubseteq G\}, \{1:A \xrightarrow{\sqsupseteq} 2:G\}\rangle$).

Conversely, if some T_i of $\mathfrak{T}$ is not satisfiable, then the entire DTB is unsatisfiable, because there is no distributed interpretation for it, and hence everything can be deduced from $\mathfrak{T}$, because d-entailment quantifies over the set of distributed interpretations, which in this case is *empty*. This is an unsatisfactory state of affairs for distributed IS.

Let us consider some possibilities in the case of $\mathfrak{T}_{12}$. (i) If just T_2 is unsatisfiable, we do not want this to affect reasoning as seen in IS_1, especially because of the "no back-flow" stance. So reasoning in $\mathfrak{T}_{12}$ should reduce to reasoning in T_1. (ii) If just T_1 is unsatisfiable, there is still some desire for this inconsistency not to "infect" the reasoning of IS_2, at least not to the point that *all* conclusions of the form $2:X \sqsubseteq Y$ are entailed by $\mathfrak{T}_{12}$. (iii) To help the integrator, it would be desirable to be able to distinguish the above cases from the one where the inconsistency is due the effect of bridge rules.

We would like to find a semantic solution having these properties in order to be sure of its coherence. So let us introduce a new, special DL interpretation, $\mathcal{I}^\delta = \left\langle \Delta^{\mathcal{I}^\delta}, \cdot^{\mathcal{I}^\delta} \right\rangle$. $\Delta^{\mathcal{I}^\delta}$ is any non empty set, and $\cdot^{\mathcal{I}^\delta}$ makes the denotation of every description be the whole domain $\Delta^{\mathcal{I}^\delta}$. Intuitively $\mathcal{I}^\delta$ provides an interpretation even to a locally inconsistent T-box; indeed, in $\mathcal{I}^\delta$, every subsumption $A \sqsubseteq B$ is satisfied, including $\top \sqsubseteq \bot$. This means that even if a DTB has an inconsistent T-box T_k, it will have *some* distributed interpretations $\{\mathfrak{I}_i\}_{i \in I}$ – ones where $\mathfrak{I}_k = \mathcal{I}^\delta$, which means that in the definition of d-entailment we will not be quantifying over the empty set of distributed interpretations. Moreover, the addition of this special interpretation does not change the set of theorems and the set of logical consequences of any description logics, which therefore maintains all its formal and computational properties.

Suppose we repeat all previous definitions, but using this more encompassing notion of satisfaction, to obtain *δ-satisfies* and *δ-entails* $\models_\delta$.

The following proposition shows that when T_2 is inconsistent, we have the desired effect, and more generally, we get "no backflow" in all cases.

Proposition 5.1. $\mathfrak{T}_{12} \models_\delta 1{:}A \sqsubseteq B$ *iff* $T_1 \models 1 : A \sqsubseteq B$.

In the case when T_1 is inconsistent we have:

Proposition 5.2. *If T_1 is an inconsistent T-box, then $\mathfrak{T}_{12} = \langle T_1, T_2, \mathfrak{B}_{12}\rangle \models_\delta 2{:}C \sqsubseteq D$ iff $T_2' \models 2{:}C \sqsubseteq D$, where T_2' is obtained by extending T_2 with $\{ H \sqsubseteq G$ such that there exist A, B, and bridge rules $1{:}A \xrightarrow{\sqsubseteq} 2{:}G$ and $1{:}B \xrightarrow{\sqsupseteq} 2{:}H$.*

This shows that when T_1 is inconsistent, results in T_2 are only affected by the bridge rules. (A less ad-hoc non-standard model theory for Distributed FOL is studied in [3], and we plan to pursue it in the future.)

6 Extensions

The previous results can be generalized in a number of directions.

First, one can consider additional kinds of bridge rules. One way to focus the search for these is to try to preserve Theorem 4.1, which means that the new bridge rules should be expressible using the special roles R_{ij}. For example, we can state that

- every A-object has at least one corresponding G-object; (in the global DL this would correspond to $A \sqsubseteq\ \geqslant 1\, \mathsf{R}_{12}.G$
- that the correspondence relation from IS_1 to IS_2 is the identity: $\mathsf{R}_{12} = id(\top)$
- that the correspondence relation is functional: $\top_i \sqsubseteq \forall \mathsf{R}_{12}.(\leqslant 1\, \mathsf{R}_{12}^-)$

Second, we can investigate the properties of DDL more complex than just $\mathfrak{T}_{12} = \langle T_1, T_2, \mathfrak{B}_{12}\rangle$. In particular, given a general $\mathfrak{T} = \langle \{T_i\}_{i\in I}, \{\mathfrak{B}_{jk}\}\rangle$, we can consider when/whether reasoning in $\mathfrak{T}$ can be reduced to reasoning in a single DL or a simpler DTB.

To do so, consider a graph where each T_i is a node i, and each *nonempty* set of bridge rules $\mathfrak{B}_{jk}$ is an edge from j to k.

If this graph is finite and acyclic, one can repeatedly eliminate the "topmost" *IS* from which information flows. In the case when the graph is cyclic, the repeated application of bridge rules can be represented by an equivalent *infinite acyclic* DTB. When the set of bridge rules is finite and the "no back-flow" condition holds, there is a finite upper bound n such that no new subsumptions statements are proved in T-boxes with index larger than n, so that the above infinite DTB has an equivalent finite DTB.

A third issue is raised by the introduction of individual objects. In standard description logics, this leads to distinguishing between T-boxes, which contain information about concepts, and A-boxes, which usually contain information about individuals. Not only is this distinction relevant from a philosophical view point (definitional vs. contingent information), but it also introduces new questions (deciding whether an individual is an instance of a concept), and it can have significant computational effects: for certain DLs, the complexity of deciding concept membership is different than that of deciding concept subsumption. These distinctions collapse in those cases where the DL has constructors, such as enumerated sets whose elements are individual objects, that allow A-box information to be encoded in the form of subsumptions involving concepts, and where individual reasoning reduces to concept reasoning in the presence of background theories containing subsumptions. In this paper, we have treated individuals exactly in this manner, although in Example 2.1, the only individual in the DTB, `"Irst"`, had no roles, so that {`"Irst"`} is known to be replaceable by a primitive concept for the purposes of reasoning.

However, if component DLs do not support constructors such as set enumeration we need to make an explicit effort to add individuals. First, we need to add A-boxes describing the contents of each local information system's database of facts. Formally, this is achieved by simply adding a set $\{A_i\}_{i \in I}$ to the description of a distributed information system. Second, we need to describe *individual mappings* like the one from `couple23` to `Gianni` and `Mary`, mentioned in Example 1.1. Model theoretically, such mappings specify (portions) of the r relation for each distributed interpretation $\mathfrak{I} = \langle \{\mathcal{I}_i\}_{i \in I}, r \rangle$. If we wanted to help extend Theorem 4.1, which describes the translation of DDL reasoning into a single "global" DL reasoner, we would introduce individual mapping rules of the form $i\!:\!x \mapsto j\!:\!y$, where x and y are individuals of $DLog_i$ and $DLog_j$ respectively. For example, we could assert that `couple23` $\mapsto$ `Gianni`. Such assertions would then translate naturally into A-box assertions of the form $x \ \mathsf{R}_{ij} \ y$ for a global A-box.

However, such assertions cannot fully capture bridge rules such as
$1 : \{\mathsf{couple23}\} \xrightarrow{\sqsubseteq} 2 : \{\mathsf{Gianni,\ Mary}\}$ and $1 : \{\mathsf{couple23}\} \xrightarrow{\sqsupseteq} 2 : \{\mathsf{Gianni,\ Mary}\}$, which really assert that $r_{12}(\text{couple23}) \subseteq \{$ `Gianni,Mary`$\}$ and $r_{12}(\text{couple23}) \supseteq \{$ `Gianni,Mary`$\}$. More specifically, the $\mapsto$ rules cannot describe the precise set of individuals being mapped to. For this purpose, we may prefer to introduce individual mapping rules of the form $i : x \xrightarrow{=} \{y_1, y_2, ...\}$, which indicate that $r_{ij}(x) = \{y_1, y_2, ...\}$. Translating these into the global A-box will however require the use of additional concept constructors such as number restrictions. It is part of our plans for future work to investigate these issue.

7 Summary

The seminal work of Catarci and Lenzerini [2] on integrating ISs described by DLs, made an implicit assumption that the local IS's have the same notion of what individual objects are, and that there was only one set of (subsumption) assertions relating IS_1 and IS_2. We have argued that in cases such as federated IS, when there is no single global view, these conditions need to be relaxed, by allowing general relationships between objects in the local domains, and by having "directed" import assertions. These intuitions were formalized in DDL using the notion of bridge rules. We note that [1] can also provide directed integration rules, but these are general Horn logic clauses, for which no reasoning is supported.

We have identified other desirable properties of federated DL, such as "no feedback", and localizaling the effect of inconsistencies so that one IS does not "infect" the reasoning of the entire system. In fact, we proposed a tentative model theory which has some of these properties.

Among the interesting results obtained are a translation of DDL reasoning to DL reasoning in the presence of qualified existential restrictions and general theories, which provides decidable subsumption algorithms for cases such as the one when all local *IS*s have $\mathcal{ALCNR}$ or $\mathcal{SHIQ}$ T-boxes.

References

[1] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. A principled approach to data integration and reconciliation in data warehousing. In *Proc. Intl. Workshop on Design and Management of Data Warehouses (DMDW'99), Heidelberg, Germany*, page 16, 1999.

[2] T. Catarci and M. Lenzerini. Representing and using interschema knowledge in cooperative information systems. *International Journal of Intelligent and Cooperative Information Systems*, 2(4):375–398, 1993.

[3] C. Ghidini and L. Serafini. Distributed First Order Logics. In D. Gabbay and M. de Rijke, editors, *Frontiers Of Combining Systems 2*, Studies in Logic and Computation, pages 121–140. Research Studies Press, 1998.

[4] I. Horrocks and U. Sattler. A description logic with transitive and inverse roles and role hierarchies. *J. Logic and Computation*, 9(3):385–410, 1999.

[5] A. Pirotte, E. Zimanyi, D. Massart, and T. Yakusheva. Materialization: A Powerful and Ubiquitous Abstraction Pattern. In *Proc. Conf. on Very Large Database (VLDB'94)*, pages pp. 630–641, Morgan Kauffman, 1994.