On the Local Closed-World Assumption of Data-Sources*

Alvaro Cortés-Calabuig[1], Marc Denecker[1], Ofer Arieli[2], Bert Van Nuffelen[1], and Maurice Bruynooghe[1]

[1] Department of Computer Science, Katholieke Universiteit Leuven, Belgium
{Alvaro, Marc.Denecker, Bert.VanNuffelen, Maurice.Bruynooghe}@cs.kuleuven.ac.be
[2] Department of Computer Science, The Academic College of Tel-Aviv, Israel
oarieli@mta.ac.il

Abstract. The *Closed-World Assumption* (CWA) on a database expresses that an atom not in the database is false. The CWA is only applicable in domains where the database has complete knowledge. In many cases, for example in the context of distributed databases, a data source has only complete knowledge about part of the domain of discourse. In this paper, we introduce an expressive and intuitively appealing method of representing a *local closed-world assumption* (LCWA) of autonomous data-sources. This approach distinguishes between the data that is conveyed by a data-source and the meta-knowledge about the area in which these data is complete. The data is stored in a relational database that can be queried in the standard way, whereas the meta-knowledge about its completeness is expressed by a first order theory that can be processed by an independent reasoning system (for example a *mediator*). We consider different ways of representing our approach, relate it to other methods of representing local closed-word assumptions of data-sources, and show some useful properties of our framework which facilitate its application in real-life systems.

1 Introduction and Motivation

In recent years, information integration has attracted considerable attention from the AI and databases communities. Generally speaking, the idea is, given a set of independent data-sources, to characterize the collective knowledge represented by them in terms of a uniform vocabulary, called the global schema, and then to exploit this information to obtain correct answers from the whole system (see [8] for a detailed description of this problem, and [1,9,14] for some particular solutions for it). An important aspect of this research is to arrive at an exact description of the information endorsed by each and every data-source in the system. Typically, a data-source stores a database consisting of a set of tuples. In standard database settings, the information held by the data-source would be

* This work is supported by FWO-Vlaanderen, European Framework 5 Project WASP, and by GOA/2003/08.

C. Baral et al. (Eds.): LPNMR 2005, LNAI 3662, pp. 145–157, 2005.

expressed by the conjunction of atoms together with the *closed-world assumption* [13]. The CWA expresses the *communication agreement* that an atom that does not appear in the database is false. However, it is clear that the CWA can only be applied when the database contains complete knowledge of the domain of discourse. In a context of distributed data-sources this assumption is inherently inappropriate since a consideration of a certain data-source as a single and complete representation of the world either completely discards the other sources of information or causes contradictions among them. For this reason, some existing approaches have applied an *open-world assumption* [3,11], interpreting a data-source just as the conjunction of atoms in the database. However, we find that this does not allow to grasp more refined information that is held in distributed data systems (sometimes called mediator-based systems). We illustrate this in the following example.

Example 1. Consider a distributed traffic tax administration system, in which there is one data-source for each county, maintaining a database of car owners in that county. There is a protocol amongst the different counties so that when a car owner leaves one county $\mathcal{A}$ to live in another county $\mathcal{B}$, then county $\mathcal{A}$ immediately transfers its information to county $\mathcal{B}$, while still preserving a record of the car owner and its current status for a certain period of time, to handle all running tax demands. By the nature of the protocol, we may assume that each data-source has complete knowledge about all car owners in its county, but in general it has more information than that. Part of the tables of a particular county, say Bronx, may look as follows:

Car Owners

Name	Model	CarID
Peter Steward	Mercedes 320	Qn-5452
John Smith	Volvo 230	Bx-5242
Mary Clark	BMW 550	Bx-5462

Location

Name	Residence
Peter Steward	Queens
Mary Clark	Bronx
John Smith	Bronx

By the nature of the distributed system, this data-source has an expertise on car owners of Bronx. This meta-knowledge allows to derive further information that is not explicitly stated in the data-source, e.g. that all people that are recorded in the table Location as residents of Bronx, are actually *all* the car owners from that county. However, as the information about car owners in Queens is not complete in this data-source, one should not rely only on the tables of this source for making further conclusions about that county.

The example above shows that when the information is distributed over several independent data-sources, a different approach is needed in order to properly capture the meaning of a particular data-source. While in distributed information systems, data-sources usually have only partial knowledge about the domain of discourse, still it is often the case that a particular source is an expert about a specific area and has complete knowledge about it. We call this area the *window of expertise* of the data source. It follows that to express the *information*

held by a data source, the explicit data recorded in the data-source has to be complemented by a meta-information that describes this window of expertise. These two kinds of information should be separated as much as possible, so that one may still consider data-sources as relational databases, and process the information about their completeness by an independent reasoning system.

Following these guidelines, we represent the meta-information about the completeness of a data-source by a theory that consists of several *local closed-world assumptions* (LCWA) [4]. A LCWA refines the closed-world assumption by specifying for a certain predicate an area in which the data source contains all true tuples of the predicate. In our approach, the semantics of a LCWA is expressed by a first-order formula of a uniform syntactical form. Specifically, the contribution of this paper is the following:

- A new method for representing local closed-world assumptions is introduced. Unlike other methods for expressing such assumptions, conceived e.g. in [2,4], which are tailored for intelligent agents, our notion of LCWA is specifically devised for describing complete knowledge in relational data-sources that are part of mediator systems. This allows, in particular, to formally define the *meaning* of each and every data source in such systems.
- The representation of the local closed-world assumption considered here allows to distinguish between the explicit data of the source and the external (implicit) information about its completeness. This separation allows to query a data-source in the standard way, whereas the knowledge about its completeness can be independently processed by the mediator system.
- We present two equivalent representations of the *meaning* of data-sources. One representation is given in terms of first-order theories and the other one is based on circumscription [10] (which is the common approach for expressing LCWA in related works; see, e.g., [2]). This equivalence allows us to show how our proposal captures the intuition behind traditional approaches for LCWA, expressed in terms of higher-order languages, and how they can be reduced to first-order theories in case that certain conditions are met.

The organization of the rest of paper is the following. In Section 2 we introduce the local closed-world assumption and use it for defining the meaning of a data-source. In Section 3 we consider an alternative approach, defined in terms of second-order, pseudo-circumscriptive formulae, and show the equivalence between the two approaches. Then, in Section 4 we give some further comments and generalizations to the local closed-world assumption, and in Section 5 we discuss some other approaches to this assumption. Section 6 concludes the paper.

2 The Local Closed-World Assumption (LCWA)

Definition 1. A *data-source* S is a pair $\langle \Sigma, D \rangle$, where Σ is a vocabulary consisting of predicate symbols in a fixed relational schema $\mathcal{R}(\Sigma)$ and a finite set

$\mathcal{C}(\Sigma)$ of constants representing the elements of the domain of discourse; and D is a finite set of ground atoms expressed in terms of Σ.

Definition 2. Let $S = \langle \Sigma, D \rangle$ be a data-source and let P be a predicate that appears in D. Denote by P^S the set of tuples of P in D. We write $P(\overline{t}) \in P^S$, where $\overline{t}$ is a tuple of terms, to denote the formula $\bigvee_{\overline{a} \in P^S} (\overline{t} = \overline{a})$.

Example 2. Let $S = \langle \Sigma, D \rangle$ be the following data-source with facts about the relations $\mathsf{CarO}(\cdot,\cdot)$ (between people and their cars ID) and $\mathsf{Loc}(\cdot,\cdot)$ (between people and the place they live.)

$$\left\langle \left\{ \mathsf{CarO}/2, \mathsf{Loc}/2 \right\}, \left\{ \begin{array}{l} \mathsf{CarO}(\mathsf{JS}, \mathsf{V231}), \mathsf{CarO}(\mathsf{MC}, \mathsf{V231}), \mathsf{CarO}(\mathsf{MC}, \mathsf{B342}), \\ \mathsf{Loc}(\mathsf{JS}, \mathsf{Qn}), \mathsf{Loc}(\mathsf{MC}, \mathsf{Bx}) \end{array} \right\} \right\rangle.$$

Here, $\mathsf{Loc}^S = \{(\mathsf{JS}, \mathsf{Qn}), (\mathsf{MC}, \mathsf{Bx})\}$, hence $\mathsf{Loc}(x, y) \in \mathsf{Loc}^S$ denotes the following formula: $\big((x = \mathsf{JS}) \wedge (y = \mathsf{Qn})\big) \vee \big((x = \mathsf{MC}) \wedge (y = \mathsf{Bx})\big)$.

Standard mediator systems consist of a number of data-sources collaborating with information through a common interface, the global schema [8,14]. In such context, the data-sources can be viewed as storing information, in the form of a collection of tuples, about certain domain in the real world. However, which parts of the modeled world are accurately represented in the data-source is not recorded explicitly in the system, and so the *meaning* of the data-source remains ambiguous. With the following definition we address this problem by characterizing through a FOL expression -using the *same* language of the data-source- the cases in which the data-source contains all the valid facts. We call this the *window of expertise* of the data-source, and it is represented in the following definition by the formula Ψ.

Definition 3. A *local closed-world assumption* for a data-source $S = \langle \Sigma, D \rangle$, is a triple $\mathcal{LCWA} = \langle S, \overline{P}, \Psi \rangle$, where $\overline{P} = \{P_1(\overline{x}_1), \ldots, P_n(\overline{x}_n)\}$ is a set of atoms (the LCWA's objects) and $\Psi(\overline{y})$ (the context of the assumption) is a first-order formula over Σ with free variables $\overline{y}$ s.t. $\overline{y} \subseteq \bigcup_{i=1}^{n} \overline{x}_i$.

Note that in each $P_i(\overline{x}_i)$, the value of the variables $\overline{x}_i \cap \overline{y}$ are constrained by Ψ. For this reason we call Ψ the window of expertise, and $\exists \overline{y} \setminus \overline{x}_i(\Psi)$ the window of expertise of the predicate P_i. The intuitive meaning of the local closed-world assumption in Definition 3 is that for each $i \in \{1, \ldots, n\}$, each fact $P_i(\overline{x}_i)$ that is true in the real world and which satisfies $\exists \overline{y} \setminus \overline{x}_i(\Psi)$ should appear in the data-source.

Example 3. Let $S = \langle \Sigma, D \rangle$ the data-source of Example 2.

1. $\langle S, \{\mathsf{CarO}(x, y)\}, x = \mathsf{MC} \rangle$ intuitively indicates that the data-source S contains all true atoms of the form $\mathsf{CarO}(\mathsf{x}, \mathsf{y})$ for $x = \mathsf{MC}$.
2. $\langle S, \{\mathsf{CarO}(x, y)\}, \mathsf{Loc}(x, \mathsf{Bx}) \rangle$ indicates that S knows about all the cars of the people that live in Bx.

3. $\langle S, \{\mathsf{CarO}(x,y), \mathsf{Loc}(x,z)\}, \mathsf{Loc}(x,\mathsf{Bx})\rangle$ expresses that S contains all the data about the cars of persons living in Bx and about all people living in Bx.
4. $\langle S, \{\mathsf{CarO}(x,y), \mathsf{Loc}(x,z)\}, x = \mathsf{MC}\rangle$ indicates that S has *full knowledge* ab -out Mary Clark (i.e., a LCWA regarding everything that is concerned with MC).

Example 4. Consider the following two local closed-world assumptions:

$$\mathcal{LCWA}_A = \langle S, \{\mathsf{CarO}(x,y), \mathsf{Loc}(x,z)\}, \mathsf{CarO}(x,\mathsf{V231}) \wedge \mathsf{Loc}(x,\mathsf{Bx})\rangle$$
$$\mathcal{LCWA}_B = \langle S, \{\mathsf{CarO}(x,u), \mathsf{Loc}(y,v)\}, u = \mathsf{V231} \wedge v = \mathsf{Bx}\rangle$$

Intuitively, the difference between these two expressions is that the first one expresses a full knowledge of S about car ownership and locations of V231 owners in Bronx. On the other hand, under the second assumption, the data-source knows all people having a $V321$ including people not living in the Bronx; the data-source also knows all people living in the Bronx, including those that do not have a $V321$.

Example 5. In case of item (1) of Example 3, the local closed-world assumption may be expressed as follows:

$$\forall x\big(x = \mathsf{MC} \rightarrow \forall y(\mathsf{CarO}(x,y) \rightarrow y = \mathsf{V231} \vee y = \mathsf{B342})\big) \tag{1}$$

In case of item (2) of the same example, the local closed-world assumption may be expressed as follows:

$$\forall x\big(\mathsf{Loc}(x,\mathsf{Bx}) \rightarrow \forall y(\mathsf{CarO}(x,y) \rightarrow y = \mathsf{V231} \vee y = \mathsf{B342})\big) \tag{2}$$

These examples lead us to the following general formulation of a local closed-world assumption in terms of first-order formulae:

Definition 4. Let $\mathcal{LCWA} = \langle S, \{P_1(\overline{x}_1), \ldots, P_n(\overline{x}_n)\}, \Psi(\overline{y})\rangle$ be a local closed-world assumption for a data-source S. The formula that is *induced* from $\mathcal{LCWA}$, denoted by $\Lambda_{\mathcal{LCWA}}$, is the following:

$$\forall \overline{y}\Big(\Psi(\overline{y}) \rightarrow \forall \overline{z}\big(\bigwedge_{i=1}^{n}\big(P_i(\overline{x}_i) \rightarrow \big(P_i(\overline{x}_i) \in P_i^D\big)\big)\big)\Big)$$

where, $\overline{x} = \bigcup_{i=1}^{n} \overline{x}_i$, and $\overline{z} = \overline{x} \setminus \overline{y}$.

Note that if $\overline{P}$ is empty, then $\Lambda_{\mathcal{LCWA}}$ is tautologically true and does not specify any additional information.

Example 6. Below are, respectively, the formulae that are induced from the local closed-world assumptions of items (1) and (2) in Example 3.

1. $\forall x\big(x = \mathsf{MC} \rightarrow \forall y\big(\mathsf{CarO}(x,y) \rightarrow$
   $((x = \mathsf{JS} \wedge y = \mathsf{V231}) \vee (x = \mathsf{MC} \wedge y = \mathsf{V231}) \vee (x = \mathsf{MC} \wedge y = \mathsf{B342}))\ \big)\big)$

2. $\forall x\big(\mathsf{Loc}(x,\mathsf{Bx}) \rightarrow \forall y(\mathsf{CarO}(x,y) \rightarrow$
$((x=\mathsf{JS}\wedge y=\mathsf{V231})\vee(x=\mathsf{MC}\wedge y=\mathsf{V231})\vee(x=\mathsf{MC}\wedge y=\mathsf{B342}))\ \big)\big)$

Note that under the unique name assumption (see Note 1 below), these formulae are equivalent with those of Example 5.[1]

Definition 5. For a data-source $S=\langle\Sigma,D\rangle$, denote: $\mathfrak{D}(S)=\bigwedge_{d\in D} d$.

Now we are ready to define the meaning of a data-source (in the context of mediator systems):

Definition 6. Let $S=\langle\Sigma,D\rangle$ be a data-source and let $\mathcal{LCWA}^j=\langle S,\overline{P}^j,\Psi^j\rangle$, $j=1,\ldots,m$, be all the local closed-world assumptions of S. Then the *meaning* of S is given by the following formula:

$$\mathfrak{M}(S)=\mathfrak{D}(S)\wedge\bigwedge_{j=1}^{m}\Lambda_{\mathcal{LCWA}^j}.$$

Note 1. When $S=\langle\Sigma,D\rangle$ is the only data-source, the following two conditions are usually assumed:

- *Domain Closure Axiom*: $\mathsf{DCA}(S)=\forall x(\bigvee_{i=1}^{n} x=C_i)$
- *Unique Name Axiom*: $\mathsf{UNA}(S)=\bigwedge_{1\leqslant i<j\leqslant n} C_i\neq C_j$

where $C_1,\ldots,C_n$ are all constants in Σ. In such cases, $\mathsf{DCA}(S)$ and $\mathsf{UNA}(S)$ appear as two additional conjuncts in $\mathfrak{M}(S)$. We denote the meaning of S by $\mathfrak{M}_D(S)$, $\mathfrak{M}_U(S)$, or $\mathfrak{M}_{DU}(S)$, when the first, the second or both assumptions are imposed, respectively.

The meaning of a data-source can be understood as a first-order theory representing incomplete knowledge about the real world. In the general case this theory will be incomplete, so there will exist more than one model, the actual world corresponding to one of them. Consequently, the meaning of a data-source is not be interpreted with respect to its database but with respect to the real world.

Given a formula Ψ, denote by $\exists|_{\overline{x}}\Psi$ the existential quantification of all free variables in Ψ, except those in $\overline{x}$.

The next proposition shows that the formula $\Lambda_{\mathcal{LCWA}}$ formalizes the intuitive meaning of the local closed-world assumption $\langle S,\overline{P},\Psi\rangle$, as specified in the paragraph below Definition 3.

Proposition 1. *For $S=\langle\Sigma,D\rangle$, let $\mathcal{LCWA}=\langle S,\{P_1(\overline{x}_1),\ldots,P_n(\overline{x}_n)\},\Psi\rangle$ and $\mathcal{LCWA}_i=\langle S,\{P_i(\overline{x_i})\},\exists|_{\overline{x}_i}\Psi\rangle$ $i=1,\ldots n$. Then:*

$$\Lambda_{\mathcal{LCWA}}\equiv\bigwedge_{i=1}^{n}\Lambda_{\mathcal{LCWA}_i}$$

[1] Consider, for instance, the first formula. It is of the form $\forall x\Phi$, and the formula in Example 3–(1) is of the form $\forall x\Phi'$. For every x other than MC both Φ and Φ' are trivially true, and for $x=\mathsf{MC}$, both Φ and Φ' hold only if there is no $c\notin\{\mathsf{V231},\mathsf{B342}\}$ s.t. $\mathsf{CarO}(\mathsf{MC},c)$ is true.

Proof. The equivalence is obtained by applying some simple rewriting rules on the relevant formulae. Indeed, denote $\overline{x} = \cup_{i=1}^{n} \overline{x}_i$ and $\overline{z} = \overline{x} \setminus \overline{y}$. Then:

$$\begin{aligned}
\Lambda_{\mathcal{LCWA}} &\equiv \forall\overline{y}\big(\Psi(\overline{y}) \rightarrow \big(\forall\overline{z}\big(\textstyle\bigwedge_{i=1}^{n}\big(P_i(\overline{x}_i) \rightarrow \big(P_i(\overline{x}_i) \in P_i^D\big)\big)\big)\big)\big)\\
&\equiv \forall\overline{y}\big(\Psi(\overline{y}) \rightarrow \big(\textstyle\bigwedge_{i=1}^{n}\forall\overline{z}\big(P_i(\overline{x}_i) \rightarrow \big(P_i(\overline{x}_i) \in P_i^D\big)\big)\big)\big)\\
&\equiv \forall\overline{y}\big(\textstyle\bigwedge_{i=1}^{n}\big(\Psi(\overline{y}) \rightarrow \forall(\overline{x}_i \setminus \overline{y})\big(P_i(\overline{x}_i) \rightarrow \big(P_i(\overline{x}_i) \in P_i^D\big)\big)\big)\big)\\
&\equiv \textstyle\bigwedge_{i=1}^{n}\forall\overline{y}\big(\Psi(\overline{y}) \rightarrow \forall(\overline{x}_i \setminus \overline{y})\big(P_i(\overline{x}_i) \rightarrow \big(P_i(\overline{x}_i) \in P_i^D\big)\big)\big)\\
&\equiv \textstyle\bigwedge_{i=1}^{n}\forall(\overline{y} \cap \overline{x}_i)\big(\exists|_{\overline{x}}\Psi(\overline{y}) \rightarrow \forall(\overline{x_i} \setminus \overline{y})\big(P_i(\overline{x}_i) \rightarrow \big(P_i(\overline{x}_i) \in P_i^D\big)\big)\big)\\
&\equiv \textstyle\bigwedge_{i=1}^{n} \Lambda_{\mathcal{LCWA}_i}.
\end{aligned}$$

Thus the equivalence is obtained. □

Example 7. The assumption $\mathcal{LCWA} = \langle S, \{\mathsf{CarO}(x,y), \mathsf{Loc}(x,z)\}, x = \mathsf{MC}\rangle$, given in Example 3-(4), which says that S has full knowledge about Mary Clark, may also be represented in a modular way by the following two expressions:

$$\begin{aligned}
\mathcal{LCWA}_A &= \langle S, \{\mathsf{CarO}(x,y)\}, x = \mathsf{MC}\rangle\\
\mathcal{LCWA}_B &= \langle S, \{\mathsf{Loc}(x,z)\}, x = \mathsf{MC}\rangle
\end{aligned}$$

We say that the meaning of a data-source is consistent if it has at least one model in the standard model-theoretic sense.

Proposition 2. *Every data-source has a consistent meaning.*

Proof. We consider the case where the meaning of a data-source $S = \langle \Sigma, D\rangle$ is given by $\mathfrak{M}(S)$. The proofs for $\mathfrak{M}_D(S)$, $\mathfrak{M}_U(S)$, and $\mathfrak{M}_{DU}(S)$ (i.e., when any combination of DCA and UNA is also assumed) are similar.

Let $\mathcal{LCWA}^j = \langle S, \overline{P}^j, \Psi^j\rangle$, $j=1,\ldots,m$ be all the local closed-world assumptions for S. Then $\mathfrak{M}(S) = \bigwedge_{A\in D} A \wedge \bigwedge_{j=1}^{m} \Lambda_{\mathcal{LCWA}^j}$. To show the proposition we define an interpretation I for Σ and show that it is a model of $\mathfrak{M}(S)$. Let I be the Herbrand interpretation associated with the database of S: the domain is $\mathcal{C}(\Sigma)$ and $P^I = P^S$. By construction of I, $I \models P_i(d_{i_1},\ldots,d_{i_k})$ for every ground atom in D. When $\Psi^j(\overline{y}_j)$ is false in I, then trivially $I \models \bigwedge_{j=1}^{m} \Lambda_{\mathcal{LCWA}^j}$. When $\Psi^j(\overline{y}_j)$ true in I, then by its construction, whenever $I \models \overline{P}^j$, also $I \models P^j(\overline{x}) \in P^S$. □

The next proposition implies that for single data-sources our semantics of the local closed-world assumption is a conservative extension of Reiter's closed-world assumption for relational databases; the present approach allows to express in such cases that a single data-source has complete knowledge about the world.

Proposition 3. *Let $S = \langle \Sigma, D\rangle$ be the only data-source and let $\mathcal{LCWA} = \langle S, \overline{P}, \mathsf{TRUE}\rangle$, where $\overline{P} = \{P_1(\overline{x}_1),\ldots,P_n(\overline{x}_n)\}$ are all the predicates occurring in Σ. Then $\mathfrak{M}_{DU}(S)$ coincides with Reiter's axiomatization of the closed-world assumption* [13] *for S.*

Proof. Reiter's axiomatization of closed-world assumption of S is a first-order theory Γ, consisting of the following formulae: (1) $\mathsf{DCA}(S)$, (2) $\mathsf{UNA}(S)$, (3) the ground atomic facts in D, and (4) completion axioms for each predicate of S: $\forall\overline{x}(P_i(\overline{x}_i) \rightarrow P(\overline{x}_i) \in P_i^D)$, $i = 1, \ldots, n$.

By Definition 6, $\mathfrak{M}_{DU}(S)$ includes (1), (2) and (3), so it remains to show that $\mathcal{LCWA} = \langle S, \overline{P}, \mathsf{TRUE}\rangle$ is equivalent to (4). Indeed, the formula that is induced from this assumption is

$$\forall\overline{x}\big(\bigwedge_{i=1}^{n}\big(P_i(\overline{x}_i) \rightarrow \big(P_i(\overline{x}_i) \in P_i^D\big)\big)\big),$$

which is equivalent to the conjunction of the formulae in (4). □

3 A Circumscriptive Approach to the LCWA

In this section we consider an alternative approach to the representation of the closed-word assumption, this time by second-order formulas, and show its equivalence to the approach given in the previous section.

Consider again item 1 of Example 3. The local closed-world assumption in this case could be defined also in terms of sets as follows:

$$\{y \mid \mathsf{CarO}(\mathsf{MC}, y)\} = \{y \mid \mathsf{CarO}(\mathsf{MC}, y) \in D\}.$$

Since the set on the left-hand side of this equation is always a superset of the set on the right-hand side, the condition could be rephrased as follows:

$$\{y \mid \mathsf{CarO}(\mathsf{MC}, y)\} \subseteq \{y \mid \mathsf{CarO}(\mathsf{MC}, y) \in D\}.$$

This condition is specified in terms of a set inclusion property, and it is common to express such conditions by means of circumscriptive formulae. These formulae express the aspiration that the set of tuples of a certain predicate, satisfying a certain condition, should be as minimal as possible. It is not surprising, therefore, that a variant of the notion of local closed-world assumption presented here has already been expressed in term of circumscriptive axioms (see [2] and Section 5).

Definition 7. Let $\mathcal{LCWA} = \langle S, \{P_1(\overline{x}_1), \ldots, P_n(\overline{x}_n)\}, \Psi(\overline{y})\rangle$ be a local closed-world assumption for a data-source $S = \langle \Sigma, D\rangle$. The *pseudo-circumscriptive form* of $\mathcal{LCWA}$ is the following (second-order) formula, denoted $C(S)$:

$$\forall\overline{\Theta}\left(\mathfrak{D}(S)[\overline{P}/\overline{\Theta}] \rightarrow \left(\forall\overline{y}\left(\Psi(\overline{y}) \rightarrow \forall\overline{z}(\overline{\Theta} \leqslant \overline{P})\right) \rightarrow \forall\overline{y}\left(\Psi(\overline{y}) \rightarrow \forall\overline{z}(\overline{P} \leqslant \overline{\Theta})\right)\right)\right),$$

where $\overline{x} = \bigcup_{i=1}^{n}\overline{x}_i$, $\overline{z} = \overline{x} \setminus \overline{y}$, and

- $\overline{P} = \{P_1(\overline{x}_1), \ldots, P_n(\overline{x}_n)\}$, $\overline{\Theta} = \{\Theta_1(\overline{x}_1), \ldots, \Theta_n(\overline{x}_n)\}$, and each $\Theta_i(\overline{x}_i)$ is a predicate variable with the same arity of $P_i(\overline{x}_i)$,
- $\overline{P} \leqslant \overline{Q}$ is an abbreviation for $\bigwedge_{i=1}^{n}(P_i(\overline{x}_i) \rightarrow Q_i(\overline{x}_i))$. [2]

[2] $C(S)$ is called pseudo-circumscriptive since it differs from a pure circumscription schema by introducing the first-order formula Ψ into the representation. Just as in Definition 4, Ψ represents the context in which $\overline{P}$ should be minimal.

Definition 8. Let $S = \langle \Sigma, D \rangle$ be a data-source and let $C^j(S)$, $j = 1, \ldots, m$ be the pseudo-circumscriptive forms of its local closed-world assumptions. Denote:

$$\mathfrak{C}(S) = \mathfrak{D}(S) \wedge \bigwedge_{j=1}^{m} C^j(S).$$

Theorem 1. *For every data-source S, $\mathfrak{M}(S)$ is equivalent to $\mathfrak{C}(S)$.*

Proof. We prove the theorem for the case that $\overline{P}$ and $\overline{\Theta}$ are singletons, and that $m = 1$. The proof can be easily extended to the general case. We have to show that when $\mathfrak{D}(S)$ holds,

$$\forall \overline{y}\Big(\Psi(\overline{y}) \rightarrow \forall \overline{z}\big(P(\overline{x}) \rightarrow \big(P(\overline{x}) \in P^S\big)\big)\Big) \tag{3}$$

is equivalent to

$$\forall \Theta\Big(\underbrace{\mathfrak{D}(S)\big[P/\Theta\big]}_{(a)} \rightarrow \Big(\underbrace{\forall \overline{y}\,\Big(\Psi(\overline{y}) \rightarrow \forall \overline{z}\big(\Theta \leqslant P\big)\Big)}_{(b)} \rightarrow \underbrace{\forall \overline{y}\,\Big(\Psi(\overline{y}) \rightarrow \forall \overline{z}(P \leqslant \Theta)\Big)}_{(c)}\Big)\Big), \tag{4}$$

where, in both cases, $\overline{z} = \overline{x} \setminus \overline{y}$. Indeed,

($\Rightarrow$) Let I be a model of $\mathfrak{D}(S)$ and (3), and consider some value Θ^I in I for the predicate variable Θ. We show that if $\mathfrak{D}(S)[P/\Theta]$ is satisfied, so is the sub-formula (c) of (4), and hence the whole formula (4) is true as well. Let us prove, then, that sub-formula (c) holds. Assume that for some $\overline{y}$, $\Psi(\overline{y})$ is true in I and for some $\overline{z}$, $P(\overline{x})$ is true in I. As I is a model of (3), this implies that $P(\overline{x}) \in P^S$, i.e. for some tuple of terms $\overline{c}$ in the table of P in S, the equality $\overline{x} = \overline{c}^I$ holds in I. Since Θ^I satisfies $\mathfrak{D}(S)[P/\Theta]$, it follows that $\overline{x} \in \Theta^I$.

($\Leftarrow$) Let I be a model of $\mathfrak{D}(S)$ and (4). From $\mathfrak{D}(S)$ it follows that $\Theta \leqslant P$. It is obvious that $\Theta \leqslant P$ implies (b). Consequently (c) holds. Assume that there exist values $\overline{x}$ such that $\Psi(\overline{y})$ and $P(\overline{x})$ hold in I. To prove (3) we need to show that $P(\overline{x}) \in P^S$; or equivalently that there exists $\overline{c} \in P^S$ s.t. $\overline{x} = \overline{c}^I$. Because of (c) holds, it follows that $\overline{x} \in \Theta^I$. By our choice of Θ^I, this mean that for $\overline{c} \in P^S$, $\overline{x} = \overline{c}^I$. □

By the last theorem, the counterparts of Propositions 1, 2, and 3 in terms of $\mathfrak{C}(S)$ are also obtained.

Note 2. It is important to note that unless the data-sources consist of sets of facts, the first-order approach and the circumscriptive approach to the LCWA do *not* coincide. To see this, consider $S = \langle \{P/1\}, \{P(a) \vee P(b)\} \rangle$, and the assumption $\mathcal{LCWA} = \langle S, \{P(x)\}, \mathsf{TRUE} \rangle$. The formula in Definition 7 expresses a set inclusion minimization, and in this case it states an unconditional minimization of any extension of P. That is, an interpretation that satisfies both the disjunctive expression in S and the circumscriptive form of $\mathcal{LCWA}$, will necessarily state that either $P(a)$ or $P(b)$ is true, *but not both*. Intuitively, this can be read as "although the data-source is not complete with respect to P, at least it knows that no other element of the domain can belong to P, except of a or b (where the 'or' here is interpreted exclusively)".

4 Extensions and Additional Comments on the LCWA

4.1 LCWA with Several Data-Sources

An important (and intended) aspect of LCWA is applying it in a multiple-source environment. In this respect, it could be useful to specify a LCWA that addresses expertise obtained by the collective information in *several* data-sources. That is,

$$\mathcal{LCWA} = \langle \{S_1, \ldots, S_n\}, \overline{P}, \Psi \rangle. \tag{5}$$

should represent complete knowledge, shared by sources $\{S_1, \ldots, S_n\}$, in the context Ψ, about the predicates in $\overline{P}$. The induced formula $\Lambda_{\mathcal{LCWA}}$ of the assumption in (5) is obtained just as in the case of one data-source, when P^S is modified in the obvious way as follows:

Definition 9. Let $S_i = \langle \Sigma, D_i \rangle$, $i = 1, \ldots, n$ be n data-sources and let P be a predicate that appears in $\bigcup_{i=1}^{n} D_i$. Denote by $P^{\cup S_i}$ the set of tuples of P in $\bigcup_{i=1}^{n} D_i$, and abbreviate by $P(\overline{t}) \in P^{\cup S_i}$ the formula $\bigvee_{\overline{a} \in P^{\cup S_i}} (\overline{t} = \overline{a})$.

Now, the formula $\Lambda_{\mathcal{LCWA}}$ for the LCWA in (5) is defined just as the formula for one source, where $P(\overline{t}) \in P^S$ is replaced by $P(\overline{t}) \in P^{\cup S_i}$.

4.2 Complex Forms of LCWA

As local closed-world assumptions are first-order formulae, they can be used for expressing more complex assumptions about the information endorsed by the data-sources. For instance, a context (i.e., the third component) of one LCWA may be a formula that is induced by another LCWA, and so it is possible to 'compose' assumptions, and get, e.g., LCWA such as the following:

$$\mathcal{LCWA} = \left\langle S_2, \{Q(\overline{x})\}, \Lambda_{\langle S_1 \{P(\overline{x})\}, \mathsf{TRUE} \rangle} \right\rangle \tag{6}$$

Note that the formula that is induced by assumption (6) is in fact equivalent to $\Lambda_{\langle S_1, \{P(\overline{x})\}, \mathsf{TRUE} \rangle} \rightarrow \Lambda_{\langle S_2, \{Q(\overline{x})\}, \mathsf{TRUE} \rangle}$, and in general,

$$\Lambda_{\langle S_2, \{Q(\overline{x})\}, \Lambda_{\langle S_1 \{P(\overline{x})\}, \Psi \rangle} \rangle} = \Lambda_{\langle S_1, \{P(\overline{x})\}, \Psi \rangle} \rightarrow \Lambda_{\langle S_2, \{Q(\overline{x})\}, \mathsf{TRUE} \rangle}.$$

This idea also allows us to express more complicated assertions in terms of local closed-world assumptions. For instance, the following formula expresses that "either S_1 or S_2 has complete knowledge about P":

$$\left\{ \Lambda_{\langle S_1, \{P(\overline{x})\}, \mathsf{TRUE} \rangle} \vee \Lambda_{\langle S_2, \{P(\overline{x})\}, \mathsf{TRUE} \rangle} \right\}$$

Another possibility is to express that the assumptions about S_1 and S_2 are complementary, and so forth.

5 Related Works

The concept of a local closed-world assumption was first introduced in [4], in the context of knowledge bases for agents. The idea in that work was to represent a situation in which an agent has local closed-world information relative to a formula Φ and a knowledge base Γ, by a condition saying that every ground sentence that unifies with Φ either follows from Γ or is falsified by it. Formally:

$$\mathcal{LCWA}\ (\Phi) \equiv (\Gamma \models \Phi\theta) \vee (\Gamma \models \neg\Phi\theta) \text{ for all ground substitutions } \theta.$$

As we have noted above, a formal semantics for the definition of [4] in terms of second-order circumscription was proposed in [2]. The intuitive idea behind this semantics is the selection of only those models that satisfy the agent's knowledge-base and that are minimal with respect to the formulae for which the agent has complete information. We note, however, that the circumscriptive approach presented in [2] allows to minimize more predicates than those allowed by the pseudo-circumscriptive formula presented here. To see this consider, for instance $\mathcal{LCWA} = \langle S, \{P(x)\}, Q(x)\rangle$. Here, one may not know for which x, $Q(x)$ is true, and indeed the pseudo-circumscriptive formula of the $\mathcal{LCWA}$ does not affect $Q(x)$, but only $P(x)$ in the context of $Q(x)$. Suppose, then, that $P(a)$ is not in S, and we do not know whether $Q(a)$ is true, i.e. $Q(a)$ is not in S. In our approach, all we can derive is that if $Q(a)$ were true, then $P(a)$ would be false; but it is also possible that $P(a)$ is true but $Q(a)$ is false. Following the approach in [2], $P(x)$ and $Q(x)$ satisfy the data-source, but moreover, the intersection of $P(x)$ and $Q(x)$ should be minimal. In particular, if $P(b)$ is in S, but $Q(b)$ is not, then $Q(b)$ is considered false. So in this approach, also part of Q is minimized, not only P.

An alternative approach to express different levels of knowledge of a certain data-source with respect to the global domain is to label the predicates of the data-sources as "sound", "complete" or "exact" (see, for instance, [1,5,6]). We identify two main drawbacks with this approach. The first one is the loss of elegance and flexibility by the introduction of non-logical symbols to the representation. The second, more serious problem, is related to the limitation in grasping more refined knowledge about the specific areas in which the predicates of the data-source contain complete information, as observed in several examples in this paper.

In [12], the concepts of "coverage" and "density" were introduced in order to measure the completeness of data-sources at the intensional and extensional levels, respectively. The authors use these concepts to determine the completeness of one or more data-sources, gathered under merge operators. As in our approach, the intension and the contents of the predicates in a data-source are divided into two independent components. This allows to provide a general completeness measure for the data-sources, but again, it is not possible to explicitly specify situations in which the data-sources have complete knowledge about (parts of) the domain of discourse.

6 Conclusion and Future Work

In this paper we presented a method of expressing the meaning of a data-source in the context of information systems that mediate among several sources. A key issue in this respect is the ability to properly define and represent particular cases where there is a complete knowledge, although partial knowledge of the sources is usually assumed. The resulting theory is expressed by a first-order one. It may also be represented by circumscriptive-like formulae.

This is an ongoing work which is part of a larger project aiming to represent and reason with incomplete information in general mediator-based systems. In such broader context a number of related issues should be addressed as well. Below we consider some of them.

- Expressing meta-knowledge about the data-sources themselves. For instance, while it is possible to express by our approach statements such as "the data-source S contains complete knowledge about car owners in Bronx", it is not possible to represent a statement such as "for every car in Bronx that is known to the data-source S, S also knows its owners". While the first statement refers to the knowledge that S possesses about the domain of discourse, the latter expresses knowledge about S itself. In order to represent the second kind of statements, an extension based on modalities in the spirit of [7] seems to be a natural candidate.
- Consider the assumption $\mathcal{LCWA}^* = \langle S, \{P(x)\}, \neg Q(x)\rangle$. If no other assumption mentions Q in its second component, this assertion does not allow to conclude whether S has complete knowledge about P. Indeed, the induced formula in this case is of the form $\Lambda_{\mathcal{LCWA}} = \forall x.\neg Q(x) \rightarrow \ldots$, but the validity of $\neg Q(x)$ cannot be verified, since the data-sources mention only positive information. Of course, if there are other assumptions, for instance, $\mathcal{LCWA}^{**} = \langle S, \{Q(x)\}, \mathsf{TRUE}\rangle$ (which implies complete knowledge about Q) then $\mathcal{LCWA}^*$ states that for all x such that $Q(x)$ is not in the database, if $P(x)$ is true then S contains $P(x)$. This situation shows that in order to obtain complete knowledge about an arbitrary predicate P under its window of expertise, the formula Ψ must define unambiguously such window. The specific conditions for which Ψ define complete knowledge over source predicates is a crucial issue that must be investigated in the depth, since it would allow to discriminate from a set of LCWA expressions which ones are useful in practice.
- While this paper concentrates on representation forms of the closed-world-assumption and their properties, computational aspects of reasoning with these assumptions should be considered as well. Among the issues that should be addressed is the effect of the local closed-world assumptions on the complexity and decidability of the resulting theories.
- Finding a proper way to incorporate the information that the mediator system has about its data-sources with the theory that relates the different terminologies of the data-sources and the global vocabulary (called *schema mappings*). This information may also be used for splitting global queries among the sources to obtain sound and complete answers (*query planning*).

References

1. D. Calvanese, G. De Giacomo, and M. Lenzerini. Description logics for information integration. In *Computational Logic: Logic Programming and Beyond, Essays in Honour of Robert A. Kowalski, Part II*, LNCS 2408, pages 41–60. 2002.
2. P. Doherty, W. Łukaszewicz, and A. Szalas. Efficient reasoning using the local closed-world assumption. In *Proc. 9th AIMSA*, LNCS 2407, pages 49–58, 2000.
3. O Duschka, M. Genesereth, and A. Levy. Recursive query plans for data integration. *J. Logic Programming*, 43(1):49–73, 2000.
4. O. Etzioni, K. Golden, and D. Weld. Sound and efficient closed-world reasoning for planning. *Artificial Intelligence*, 89(1-2):113–148, 1997.
5. G. Grahne. Information integration and incomplete information. *IEEE Data Engineering Bulletin*, 25(3):46–52, 2002.
6. G. Grahne and A. Mendelzon. Tableau techniques for querying information sources through global schemas. In *Proc. 7th ICDT*, LNCS 1540, pages 332–347, 1999.
7. J. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, 1990.
8. M. Lenzerini. Data integration: A theoretical perspective. In *Proc. 21st PODS*, pages 233–246, 2002.
9. Rousset M. and Reynaud Ch. Knowledge representation for information integration. *Inf. Syst.*, 29(1):3–22, 2004.
10. J. McCarthy. Applications of circumscription to formalizing common sense knowledge. In V. Lifschitz, editor, *Formalizing Common Sense: Papers by John McCarthy*, pages 198–225. Ablex Publishing Corporation, 1990.
11. T. Millstein, A. Levy, and M. Friedman. Query containment for data integration systems. In *Proc. 21st PODS*, pages 67–75, 2002.
12. F. Naumann, J.C. Freytag, and U. Leser. Completeness of integrated information sources. *Information Systems*, 29(7):583–615, 2004.
13. R. Reiter. Towards a logical reconstruction of relational database theory. In *On Conceptual Modelling, Perspectives from Artificial Intelligence, Databases, and Programming Languages.*, pages 191–233, 1982.
14. B. Van Nuffelen, A. Cortés-Calabuig, M. Denecker, O. Arieli, and M. Bruynooghe. Data integration using ID-logic. In *Proc. 16th CAiSE*, LNCS 3084, pages 67–81, 2004.