A Context-based Name Resolution Approach for Semantic Schema Integration

Rosalie Barreto Belian

Supervisor: Profa. Ana Carolina Salgado

July, 2008

Universidade Federal de Pernambuco
Centro de Informática
Pós-Graduação em Ciências da Computação

Objective

- Specification and implementation of a name resolution approach for schema integration using contextual information
- Schema element names
  - Natural language words connected with their meanings
  - Meanings depending on the context to which they are related
  - Generic, domain-specific and contextual knowledge are necessary
    - various approaches use solely generic and domain-specific information
  - Are formed by Natural language words from any syntactical category
  - Are made-up by Multi-word terms
- *Context*
  - circumstantial elements that make a certain situation unique and comprehensible
  - a *context* contains metadata relating to its meaning, properties
    - its source, quality and precision, and organization

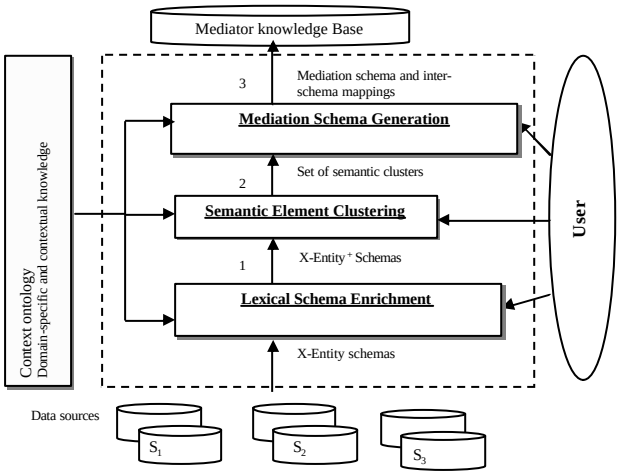
Objective

- Improve name resolution in schema integration using contextual information
  - Allowing context-bound interpretations
  - Providing richer semantic information
- To achieve this, we proposed
  - a general schema integration process
    - To validate the name resolution approach
  - an internal format to annotate schemas with lexical information
    - We are working at Integra
    - Integra uses X-Entity to represent schemas
  - the relevant contextual information for name resolution in schema integration
    - Represented as a context ontology including contextual and domain information

Name Resolution in Schema Integration

- Determine which real-world object a given schema element (entity, attribute or relationship) refers to
  - Spell-check and expansion of abbreviations and acronyms
  - Schema element sense disambiguation
  - Mediation schema element naming

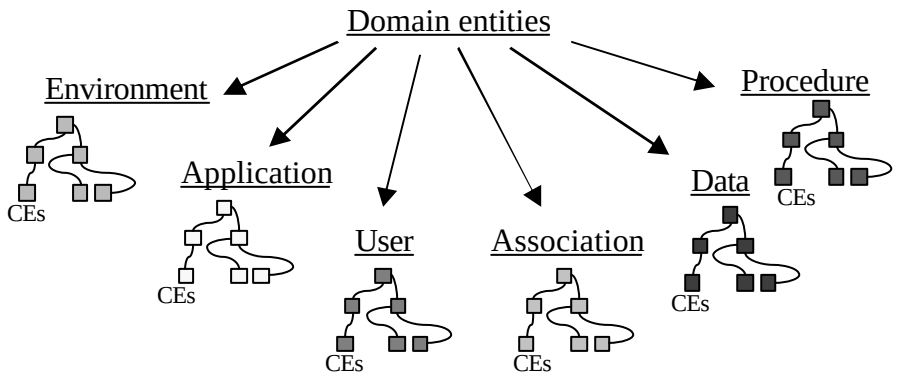
The Schema Integration Process



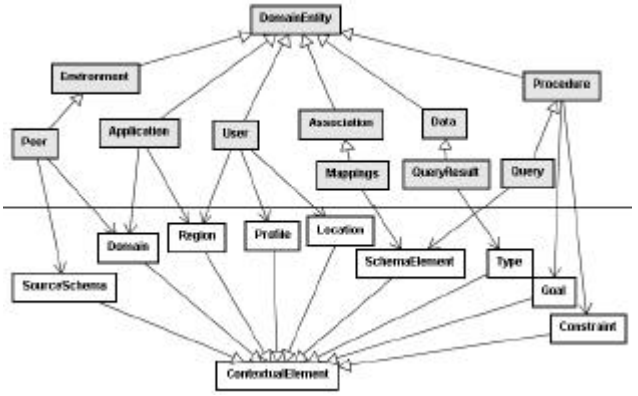
The Context Ontology

- Semantic information
  - Generic, domain-specific (UMLS)
  - Contextual information
- Contextual information necessary to schema integration/name resolution
  - Domain entities
  - Contextual elements (CEs)
- Context ontology
  - Reasoning, sharing of information
  - User-rules do declare specific situations that affect the meaning of terms such as defined in the domain
- Domain entities to DI (data integration)

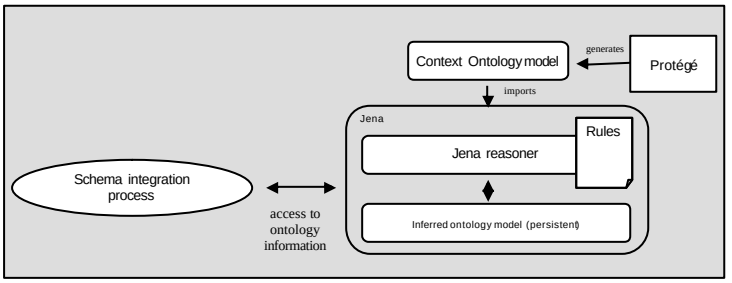
The Context Ontology



The Context Ontology



Prototype Implementation



- Context ontology
  - OWL DL using Protégé 3.2
  - Jena to reason on user-defined rules
  - W3C/SPARQL query language to access ontology information
  - UMLS 2007AC files
- Schema integration process
  - Java™ programming language version 1.5

Conclusion

- Context-based name resolution approach
  - Contextual information providing enriched semantic information for name resolution
  - Data source semantics (based on local semantics)
  - Mediation semantics (based on user and applications characteristics)
  - Ontology-based approach including domain-specific and contextual information
    - Augmenting the domain information with the contextual information required to name resolution
  - Model of relevant contextual information to name-resolution
    - May be used in other information integration issues

Future Work

- Inclusion of instance-level information in the Schema integration process
- Evaluation of the Name resolution approach (with and without context)
- Refinement of the Clustering process
- Evaluation of the generated Mediation schema
- Specification of a process to acquisition of generic and domain-specific knowledge
- Specification of a semantic similarity measure (Contextual? Feature-based?)

A Context-based Name Resolution Approach for Semantic Schema Integration

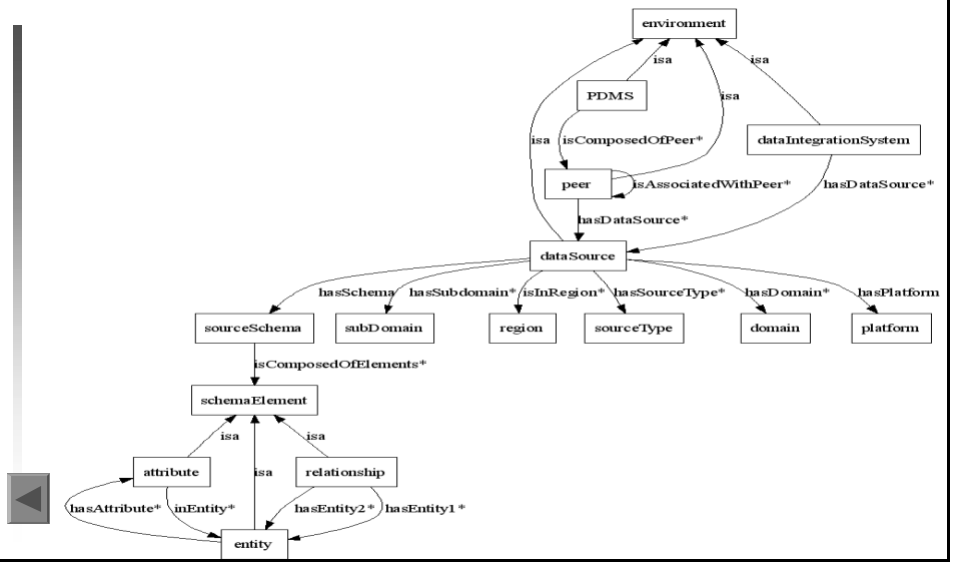
Rosalie Barreto Belian

Supervisor: Profa. Ana Carolina Salgado

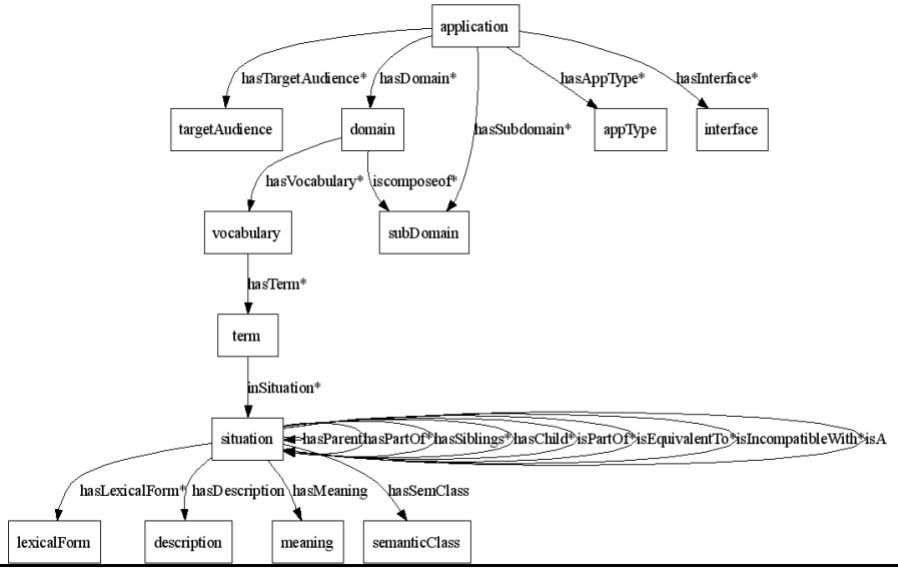
July, 2008

Universidade Federal de Pernambuco
Centro de Informática
Pós-Graduação em Ciências da Computação

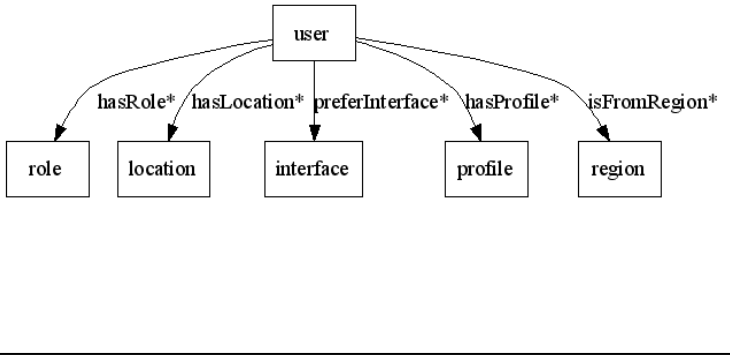
Environment CEs



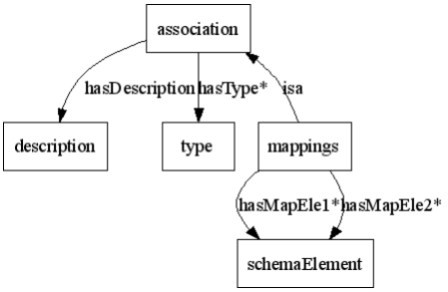
Application CEs



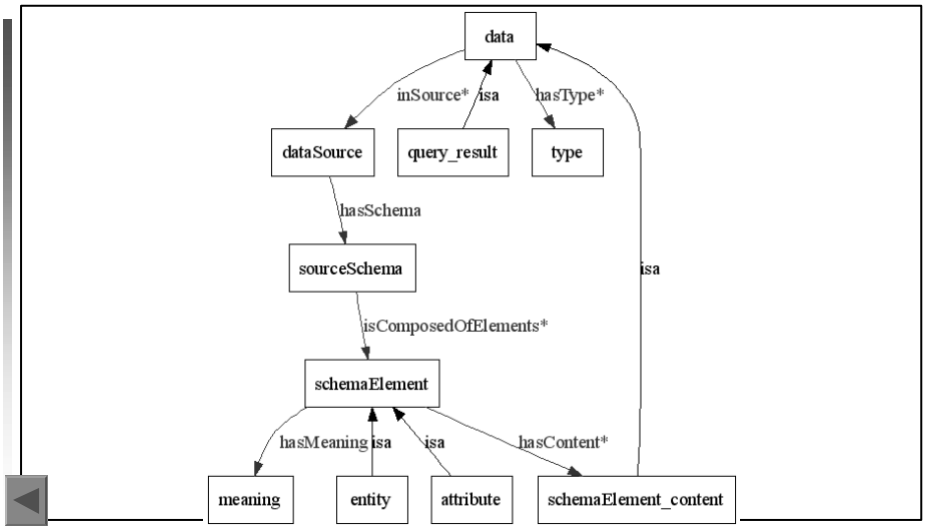
User CEs



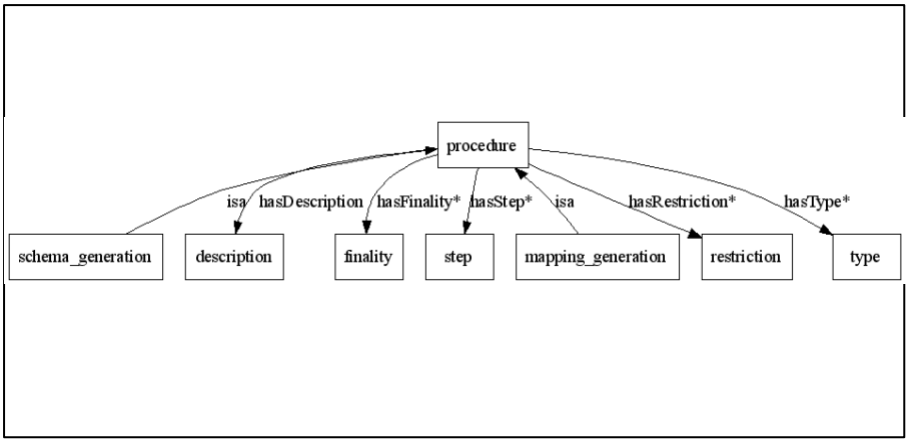
Association CEs



Data CEs



Procedure CEs



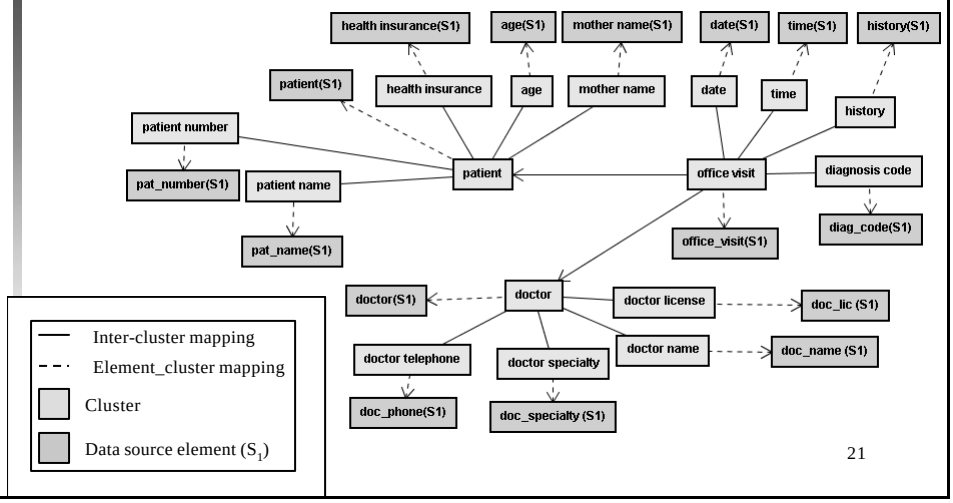
Lexical Schema Enrichment

- Step1 – preprocessing of element names from data source schemas
  - tokenization, unnecessary character removal, stop-words removal, case changes handling
- Step2 – searching schema element names in the ontology
  - Expansion of abbreviations and acronyms
  - Verification of the correct spelling
  - Element name sense disambiguation
  - Annotation of lexical information in X-Entity
- It is used syntactic matching methods and also, semantic information provided by the ontology to expand tokens and help to address syntactic issues

Semantic Element Clustering

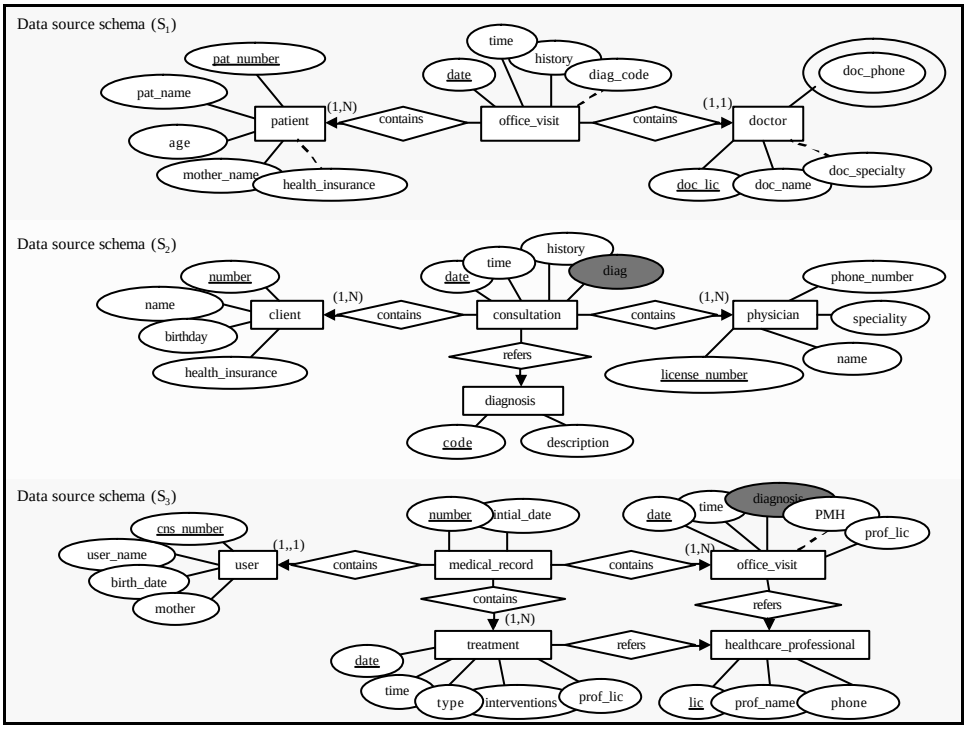
- Linguistic-based approach
  - Groups similar schema elements using their names and intended meanings
  - Semantic information is retrieved from the ontology
- Entities and attributes are grouped into the most similar related semantic cluster
- Relationships are represented as inter-cluster relationships
  - Hypernymy/hyponymy (is-a), Meronymy/holonymy (part-of), Association (refers-to) and Attribution (has).

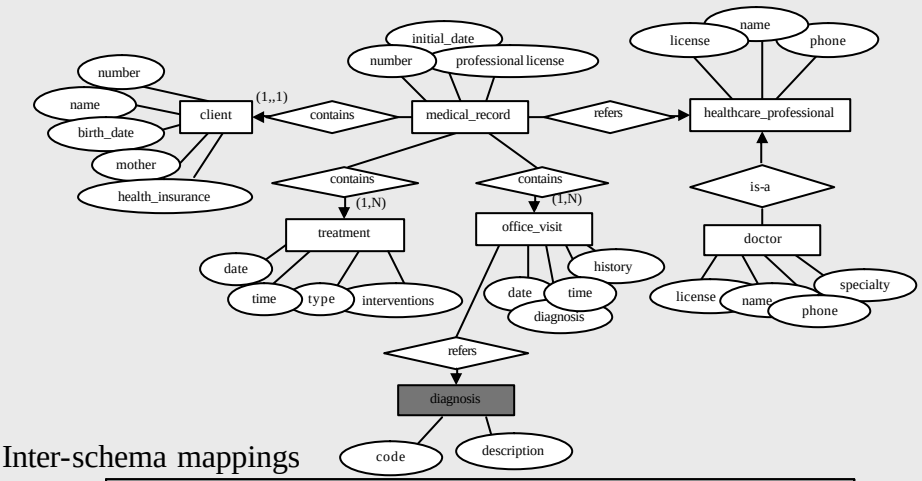
Set of Semantic Clusters (Snapshot$_1$ – S$_1$)

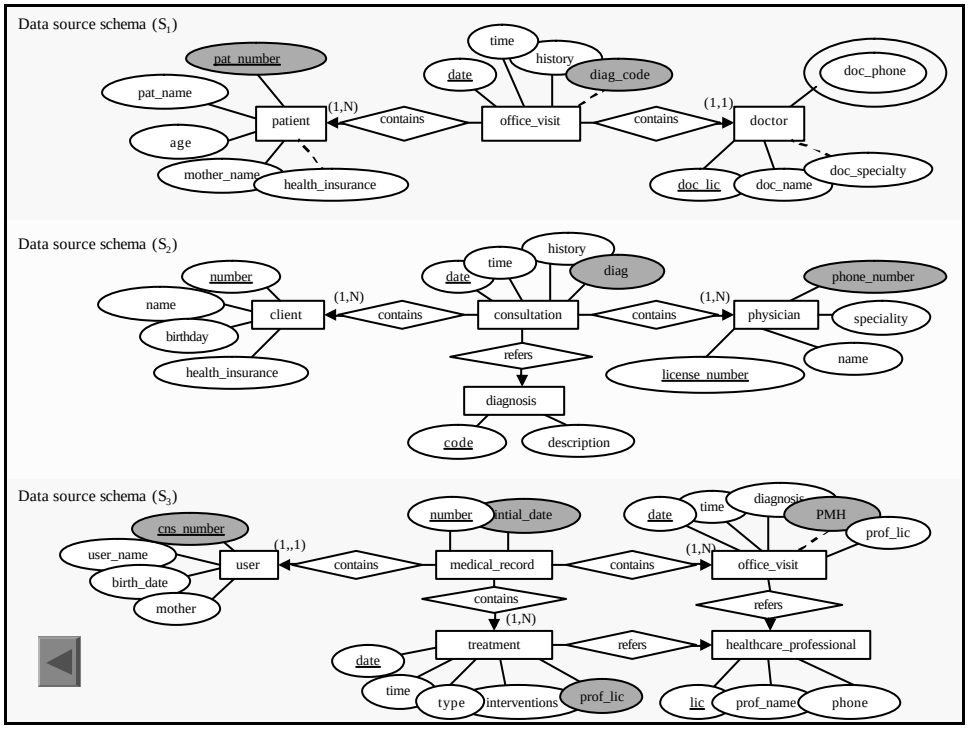


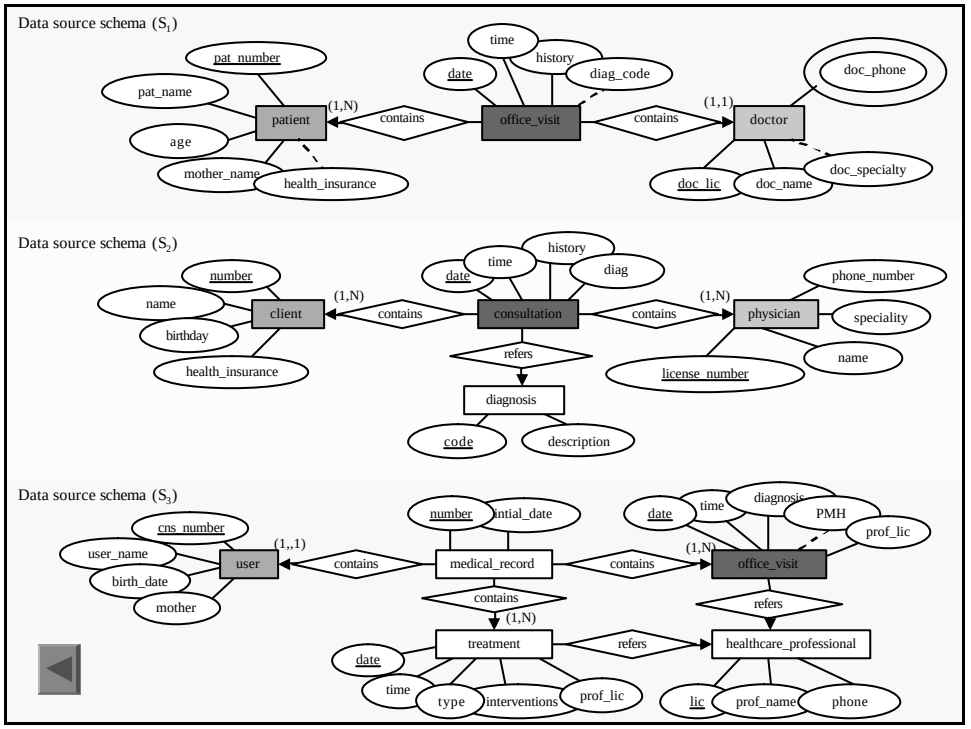
Mediation schema generation

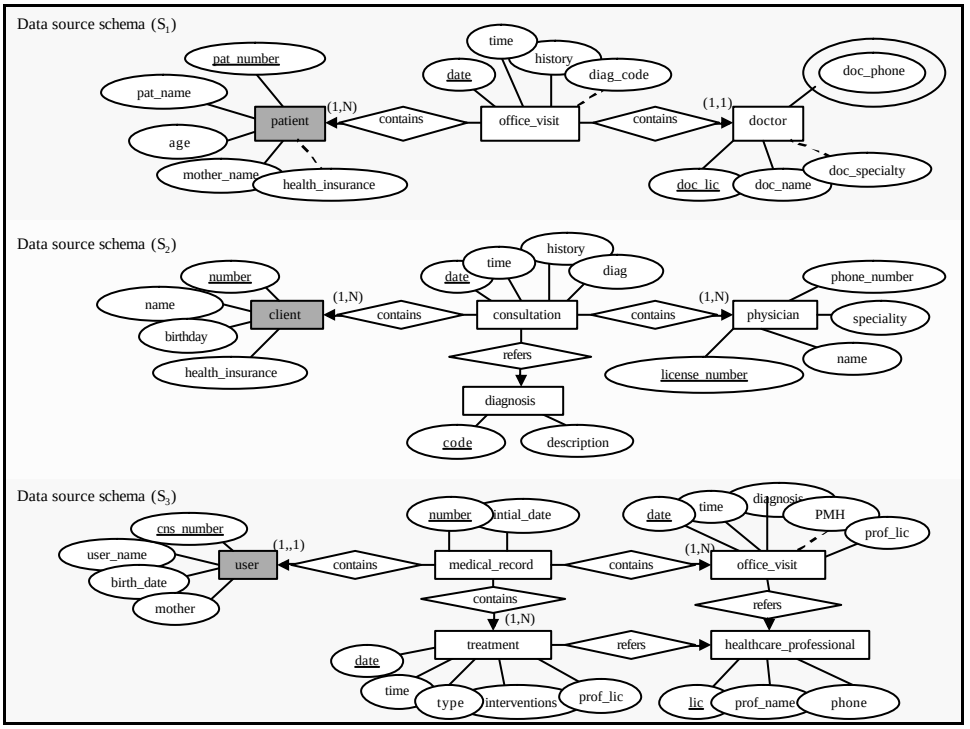
- Analyzes the set of clusters in order to
  - Generate the final mediation schema
  - Generate the inter-schema mappings (mediation vs. data source schema elements)
- Main issues
  - Definition of resulting relationships
  - Definition of inheritance relationships between mediation entities
  - Selection of final names for the mediation schema elements
  - Resolution of some structural differences
    - Mediation entity *vs.* data source attribute
    - Mediation attribute *vs*. data source entity


Data source schema (S_1)
pat_number
pat_name
age
mother_name
health_insurance
patient
(1,N)
contains
time
history
date
diag_code
office_visit
contains
(1,1)
doctor
doc_phone
doc_specialty
doc_lic
doc_name
Data source schema (S_2)
number
name
birthday
health_insurance
client
(1,N)
contains
time
history
date
diag
consultation
contains
(1,N)
physician
phone_number
speciality
name
license_number
refers
diagnosis
code
description
Data source schema (S_3)
cns_number
user_name
birth_date
mother
user
(1,,1)
contains
number
intial_date
medical_record
contains
(1,N)
date
time
diagnosis
PMH
prof_lic
office_visit
refers
contains
(1,N)
treatment
refers
healthcare_professional
date
time
type
interventions
prof_lic
lic
prof_name
phone

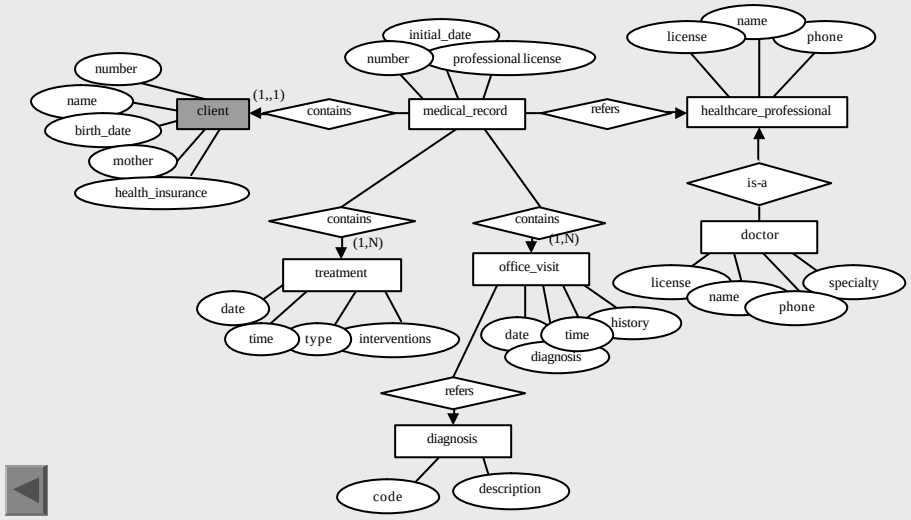

Mediation schema
number
name
birth_date
mother
health_insurance
client
(1,,1)
contains
initial_date
number
professional license
medical_record
refers
name
license
phone
healthcare_professional
contains
(1,N)
treatment
date
time
type
interventions
contains
(1,N)
office_visit
date
time
diagnosis
history
is-a
doctor
license
name
phone
specialty
refers
diagnosis
code
description
Inter-schema mappings
medical_record$_M$.medical_record$_M$_office_visit$_M$.office_visit$_M$.office_visit$_M$_diagnosis$_M$.diagnosis$_M$.code$_M$ @ office_visit$_1$.diag_code$_1$
medical_record$_M$.medical_record$_M$_office_visit$_M$.office_visit$_M$.office_visit$_M$_diagnosis$_M$.diagnosis$_M$.code$_M$ @ medical_record$_3$.medical_record$_3$_office_visit$_3$.office_visit$_3$.diagnosis$_3$


Data source schema (S_1)
pat_number
pat_name
age
mother_name
health_insurance
patient
(1,N)
contains
time
history
date
diag_code
office_visit
contains
(1,1)
doctor
doc_phone
doc_specialty
doc_lic
doc_name
Data source schema (S_2)
number
name
birthday
health_insurance
client
(1,N)
contains
time
history
date
diag
consultation
contains
(1,N)
physician
phone_number
speciality
name
license_number
refers
diagnosis
code
description
Data source schema (S_3)
cns_number
user_name
birth_date
mother
user
(1,,1)
contains
number
intial_date
medical_record
contains
date
time
diagnosis
PMH
prof_lic
office_visit
(1,N)
contains
refers
(1,N)
treatment
refers
healthcare_professional
date
time
type
interventions
prof_lic
lic
prof_name
phone


Data source schema (S_1)
pat_number
pat_name
age
mother_name
health_insurance
patient
(1,N)
contains
time
history
date
diag_code
office_visit
contains
(1,1)
doctor
doc_phone
doc_specialty
doc_lic
doc_name
Data source schema (S_2)
number
name
birthday
health_insurance
client
(1,N)
contains
time
history
date
diag
consultation
contains
(1,N)
physician
phone_number
speciality
name
license_number
refers
diagnosis
code
description
Data source schema (S_3)
cns_number
user_name
birth_date
mother
user
(1,,1)
contains
number
intial_date
medical_record
contains
date
time
diagnosis
PMH
prof_lic
office_visit
(1,N)
contains
refers
(1,N)
treatment
refers
healthcare_professional
date
time
type
interventions
prof_lic
lic
prof_name
phone

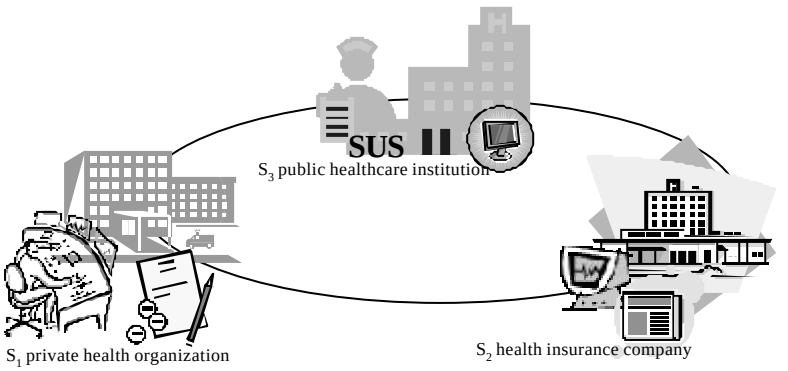


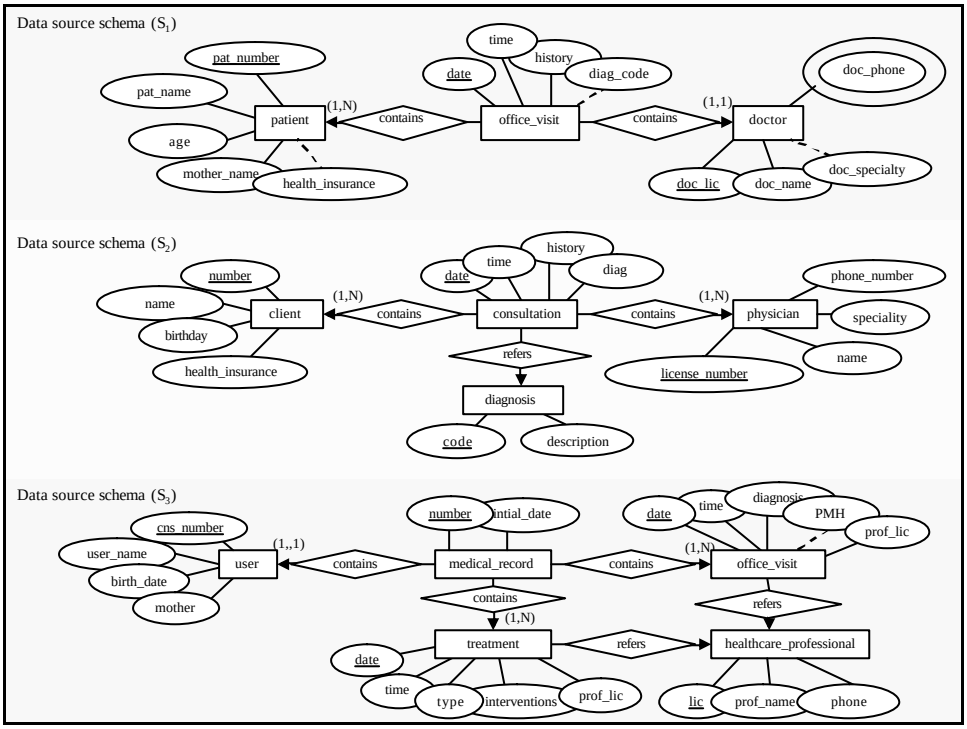
Mediation schema

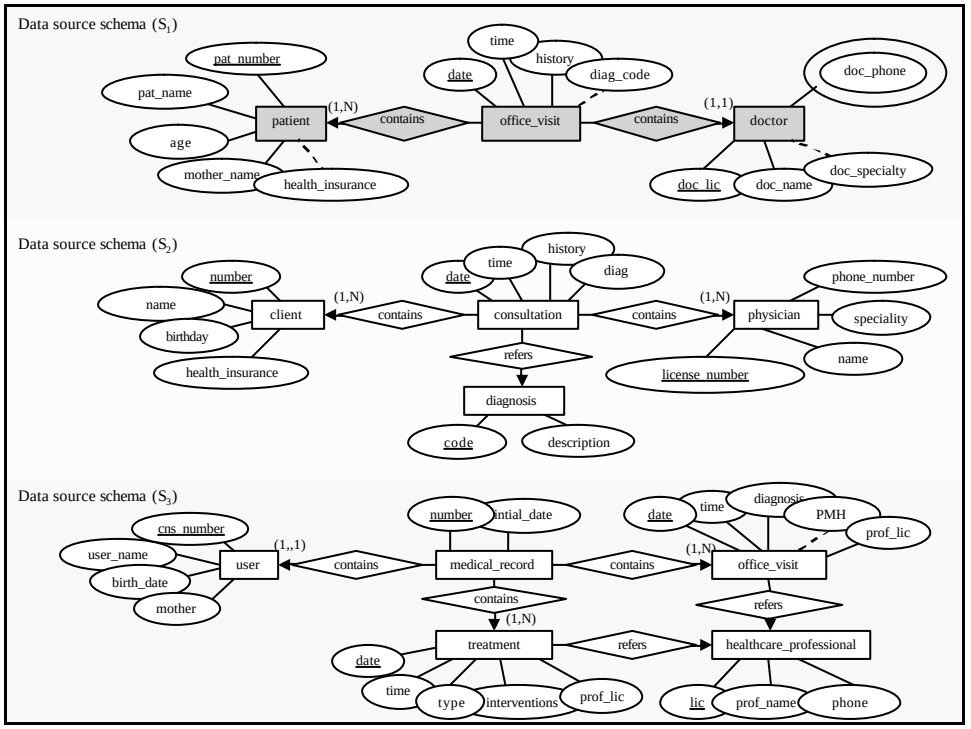


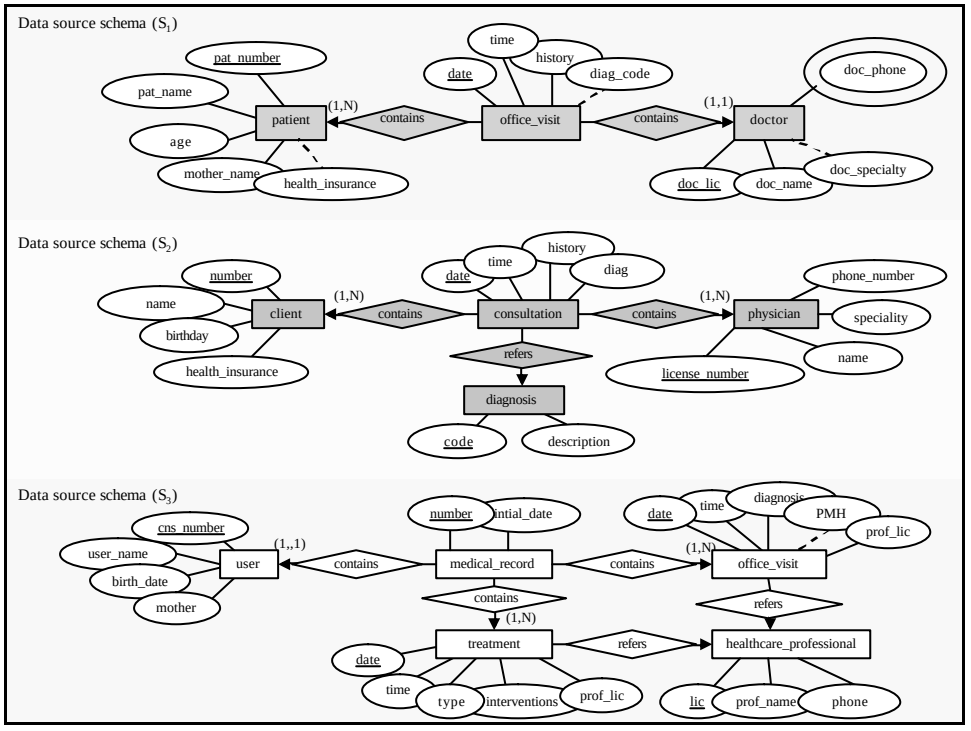
The Motivating Scenario

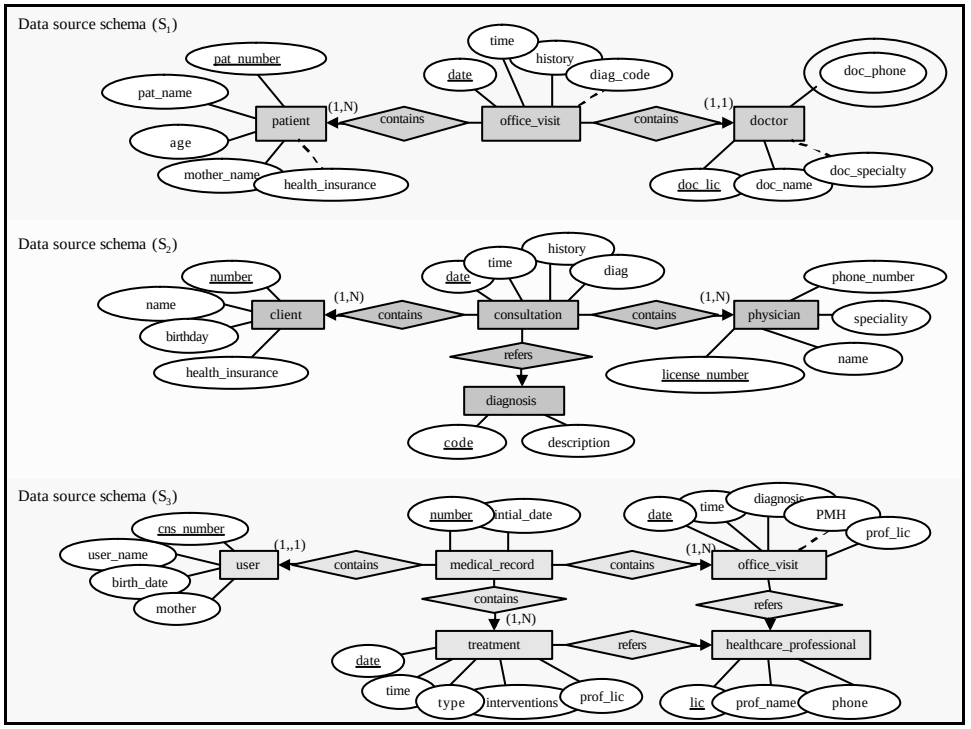
- Data sources
  - S_1: medical data from a private health organization
  - S_2: health insurance company
  - S_3: public health institution






Data source schema (S_1)
pat_number
pat_name
age
mother_name
patient
health_insurance
(1,N)
contains
time
history
date
diag_code
office_visit
contains
(1,1)
doctor
doc_phone
doc_specialty
doc_lic
doc_name
Data source schema (S_2)
number
name
birthday
health_insurance
client
(1,N)
contains
time
history
date
diag
consultation
contains
(1,N)
physician
phone_number
speciality
name
license_number
refers
diagnosis
code
description
Data source schema (S_3)
cns_number
user_name
birth_date
mother
user
(1,,1)
contains
number
intial_date
medical_record
contains
date
time
diagnosis
PMH
prof_lic
(1,N)
office_visit
contains
refers
(1,N)
treatment
refers
healthcare_professional
date
time
type
interventions
prof_lic
lic
prof_name
phone


Data source schema (S_1)
pat_number
pat_name
age
mother_name
patient
health_insurance
(1,N)
contains
time
history
date
diag_code
office_visit
contains
(1,1)
doctor
doc_phone
doc_specialty
doc_lic
doc_name
Data source schema (S_2)
number
name
birthday
health_insurance
client
(1,N)
contains
time
history
date
diag
consultation
contains
(1,N)
physician
phone_number
speciality
name
license_number
refers
diagnosis
code
description
Data source schema (S_3)
cns_number
user_name
birth_date
mother
user
(1,,1)
contains
number
intial_date
medical_record
contains
date
time
diagnosis
PMH
prof_lic
(1,N)
office_visit
contains
refers
(1,N)
treatment
refers
healthcare_professional
date
time
type
interventions
prof_lic
lic
prof_name
phone



Data source schema (S_1)
pat_number
pat_name
age
mother_name
patient
(1,N)
contains
health_insurance
time
history
date
diag_code
office_visit
contains
(1,1)
doctor
doc_phone
doc_specialty
doc_lic
doc_name
Data source schema (S_2)
number
name
birthday
health_insurance
client
(1,N)
contains
time
history
date
diag
consultation
contains
(1,N)
physician
phone_number
speciality
name
license_number
refers
diagnosis
code
description
Data source schema (S_3)
cns_number
user_name
birth_date
mother
user
(1,,1)
contains
number
intial_date
medical_record
contains
date
time
diagnosis
PMH
prof_lic
(1,N)
office_visit
contains
(1,N)
refers
treatment
refers
healthcare_professional
date
time
type
interventions
prof_lic
lic
prof_name
phone

Resulting mediation schema (S_1, S_2, S_3)