XSEM - A Conceptual Model for XML Data

Martin Necasky
martin.necasky@mff.cuni.cz, http://www.necasky.net

Intelligent Models, Algorithms, Methods and Tools for the Semantic Web Realisation
October 5 - 7, 2006

Outline

- motivation
- idea
- XSEM model
- future research

Motivation

- XML is used for/as:
  - exchange of data between heterogeneous IS
  - internal data representation
  - logical database model (pureXML, ...)
- **challenge**
  - modeling XML data should become an inseparable part of conceptual modeling of application data

Motivation
Example

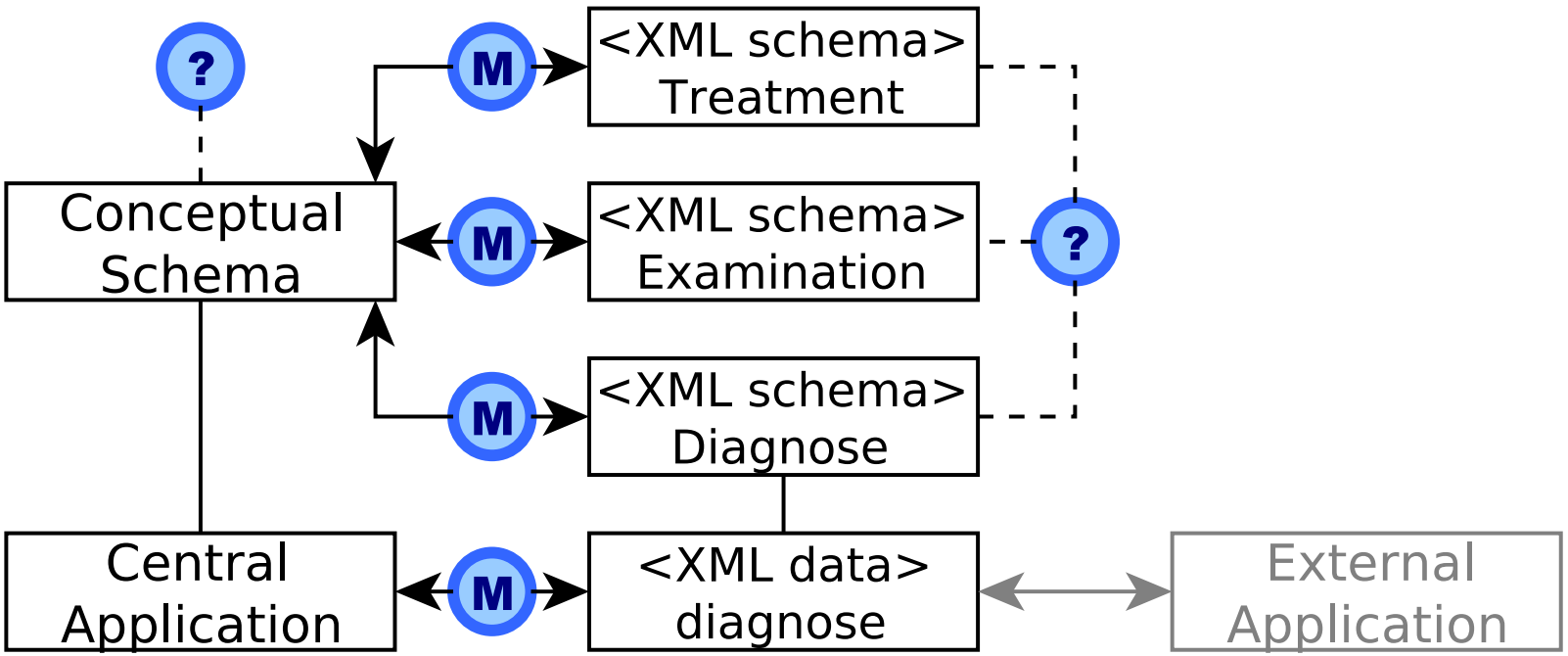
- medical application managing data about patients from several external sources
  - internal data representation
  - external sources access data through XML documents

Motivation
Example

- example scenario:
  1. physician in a hospital makes a diagnose of a patient
  2. to decide a diagnose, he or she needs results of a patient's examination
  3. hospital system requests the results from the central application
  4. central application exports the results from the internal representation into an XML document and sends it back to the hospital system
  5. physician in a hospital diagnoses a disease of a patient
  6. hospital system creates an XML document in a required form and sends it to the central application
  7. central application receives the XML diagnose and transforms it into the internal representaion

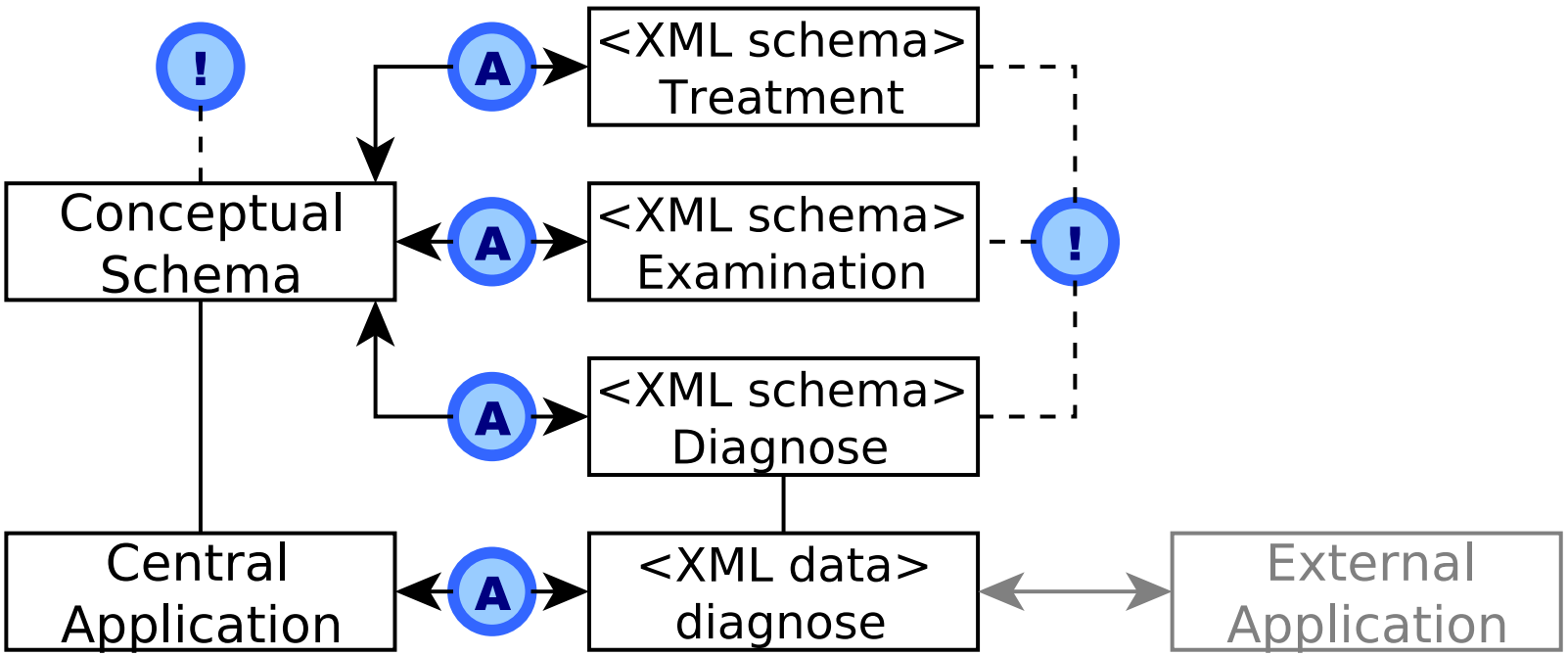
Motivation

Motivating Example - Today



Motivation

Motivating Example - Challenge



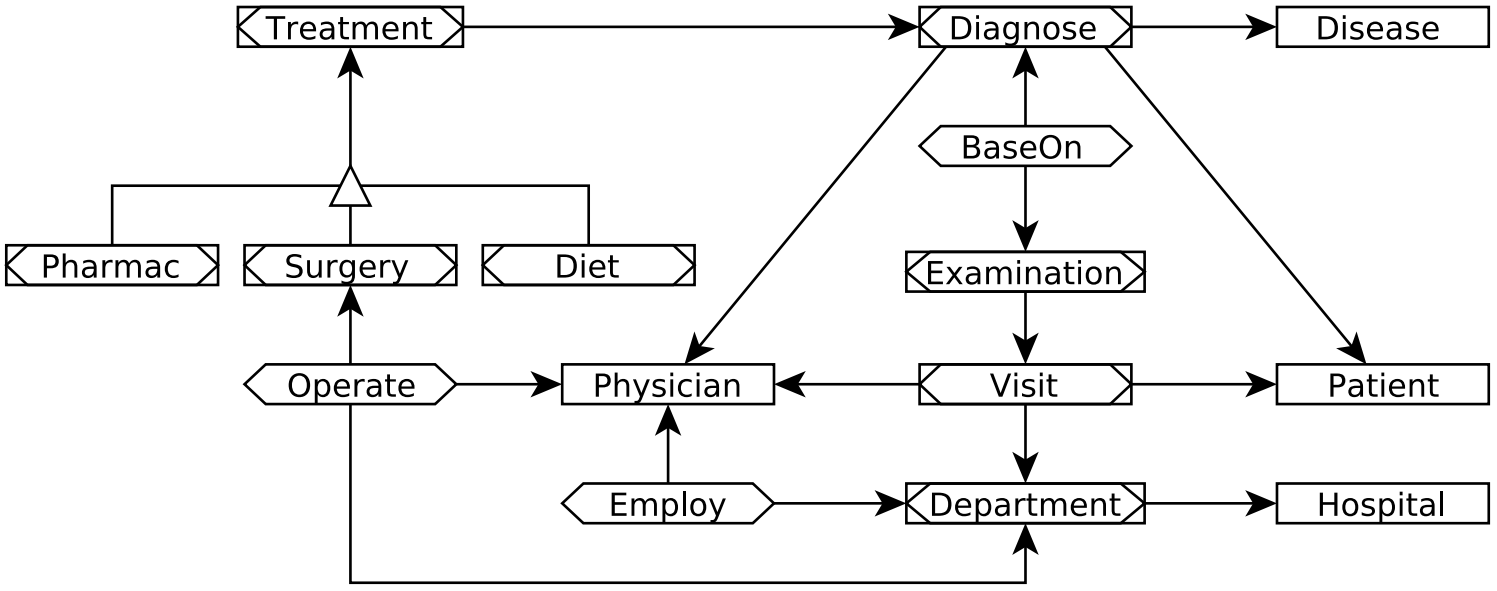
How to Model XML Data on a Conceptual Level?

XML Special Features

- irregular structure
- ordering
- structured data mixed with an unstructured text
- hierarchical

How to Model XML Data on a Conceptual Level?

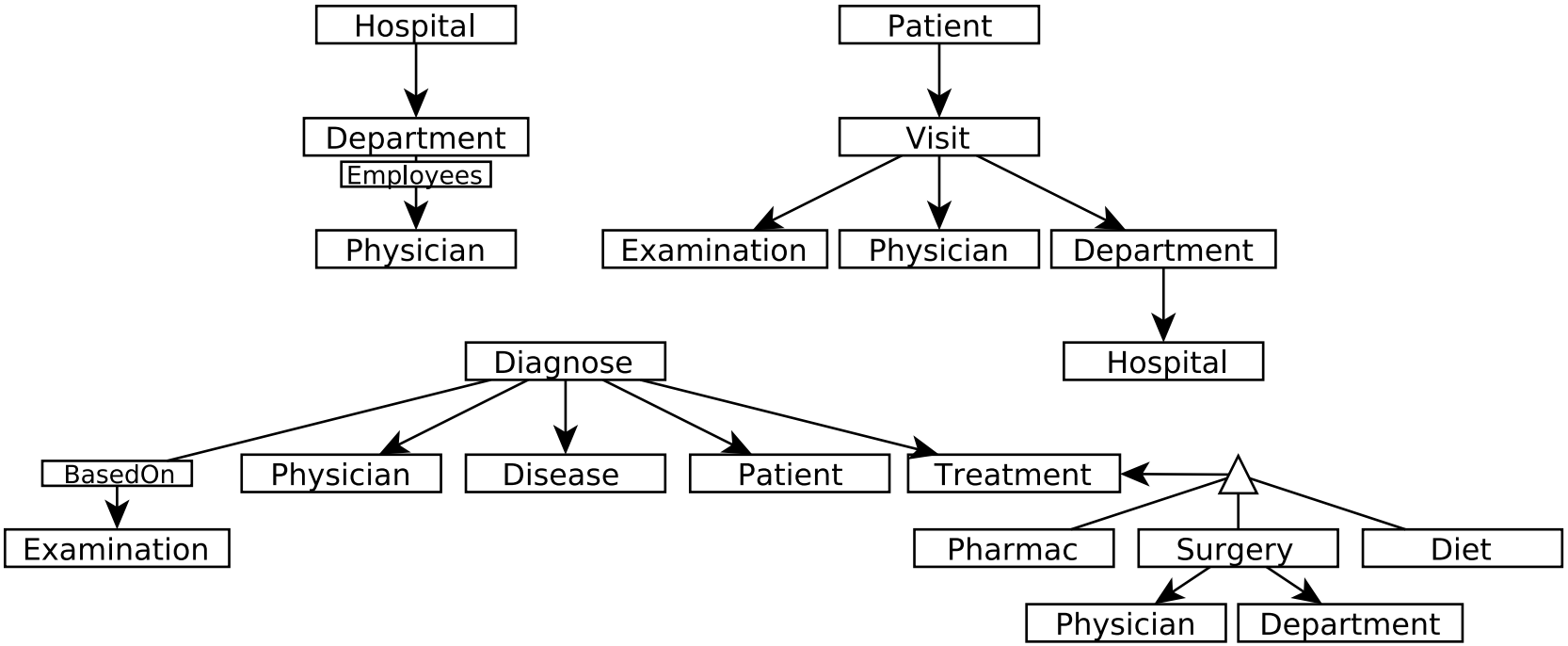
E-R Model



- there are extensions of E-R for modeling special XML features
- however, E-R is not hierarchical ($M : N$ relationship types, n-ary relationship types)
  - existing solutions derive hierarchical organization automatically without following user's requirements

How to Model XML Data on a Conceptual Level?

Modeling Hierarchical Structure



- not so clear, much more complex
- many redundancies, normalization of schemes means normalization of data on the logical level which is not desirable in every situation
- every hierarchy is only one of the possible hierarchies of the same data

Idea
Logical Level

- internal logical schema
- users access data in the internal representation through XML documents described by the XML schemes
- XML schemes as hierarchical views on parts of the internal logical schema
- each group of users needs different structure of XML documents containing the same data
  - different hierarchical views (XML schemes) on the same parts of the internal logical schema

Idea
Conceptual Level

- overall conceptual schema describing data
- hierarchical conceptual views derived from the overall conceptual schema describing required XML schemes

XSEM

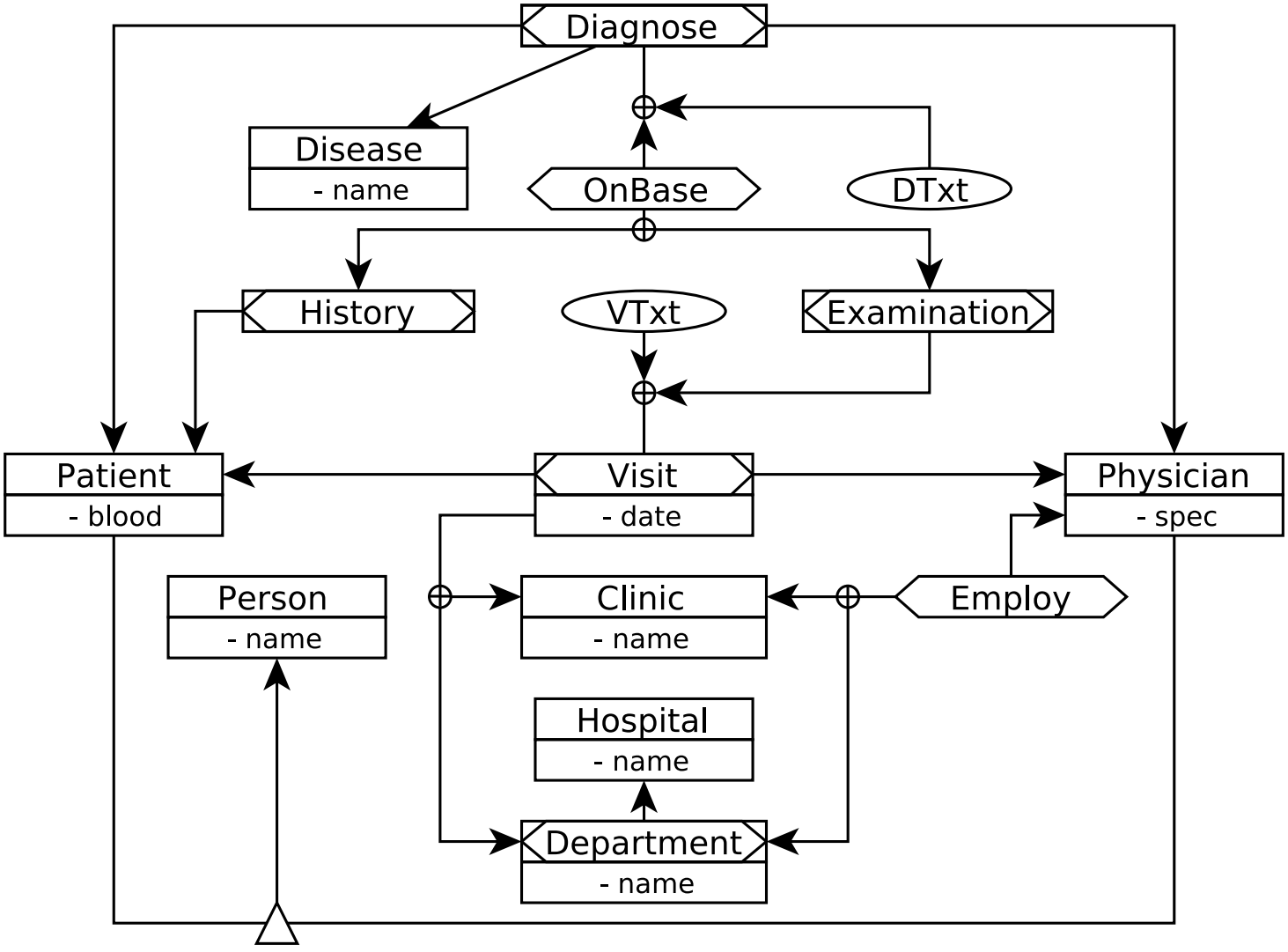
- conceptual model for XML based on the previous idea
- devides a conceptual modeling process to two parts:
  1. the first part consits of designing an overall conceptual schema of a domain using *XSEM-ER*
  2. the second part consists of designing hierarchical organizations of the structures from the first part using *XSEM-H*

XSEM-ER

- extension of E-R proposed by Chen
- modeling of special XML features:
  - irregular structure (*cluster types*)
  - ordering (*ordering constraints*)
  - mixed content (*data node types and cluster types*)
- hierarchical organization is not important here

XSEM-ER

Example



XSEM-ER
Hierarchical Projections

- step between the non-hierarchical XSEM-ER level and the hierarchical XSEM-H level
- binarization of relationship types and weak entity types
- for example, *Visit* can be represented by the following hierarchy:
  - a list of patients
  - for each patient the list of patient's visits
  - for each patient's visit the visited physician and the department or clinic where the patient visited the physisian

XSEM-ER
Hierarchical Projections

- the hierarchy describes the structure of the following example XML document:

```
<patient><name>Patient 1</name><blood>A+</blood>
   <visit><date>1.10.2006</date>
      <physician><name>Physician 1</name></physician>
      <department><name>Dept 1</name>
         <hospital><name>Hosp 1</name></hospital>
      </department>
   </visit>
   <visit><date>1.10.2006</date>
      <physician><name>Physician 2</name></physician>
      <clinic><name>Clinic 1</name></clinic>
   </visit>
</patient>
```

XSEM-ER
Hierarchical Projections

- the hierarchy is formally described by the following hierarchical projections:

$$Visit[Patient \rightarrow Visit]$$
$$Visit^{Patient}[Visit \rightarrow Physician]$$
$$Visit^{Patient}[Visit \rightarrow Department + Clinic]$$

- another hierarchy:

$$Visit[Department + Clinic \rightarrow Physician]$$
$$Visit^{Department+Clinic}[Physician \rightarrow Patient]$$
$$Visit^{Department+Clinic\ Physician}[Patient \rightarrow Visit]$$

XSEM-ER
Cardinality Constraints

- specified for hierarchical projections

$$card(Visit[Department + Clinic \rightarrow Physician], Department + Clinic) = (1, *)$$
$$card(Visit[Department + Clinic \rightarrow Physician], Physician) = (0, 5)$$
$$card(Visit^{Department+Clinic}[Physician \rightarrow Patient], Physician) = (10, 100)$$
$$card(Visit^{Department+Clinic}[Physician \rightarrow Patient], Patient) = (1, 1)$$
$$card(Visit^{Department+Clinic\ Physician}[Patient \rightarrow Visit], Patient) = (1, *)$$

XSEM-ER
Ordering

- ordering of hierarchical projections

$$order(OnBase[Diagnose \rightarrow History + Examination])$$

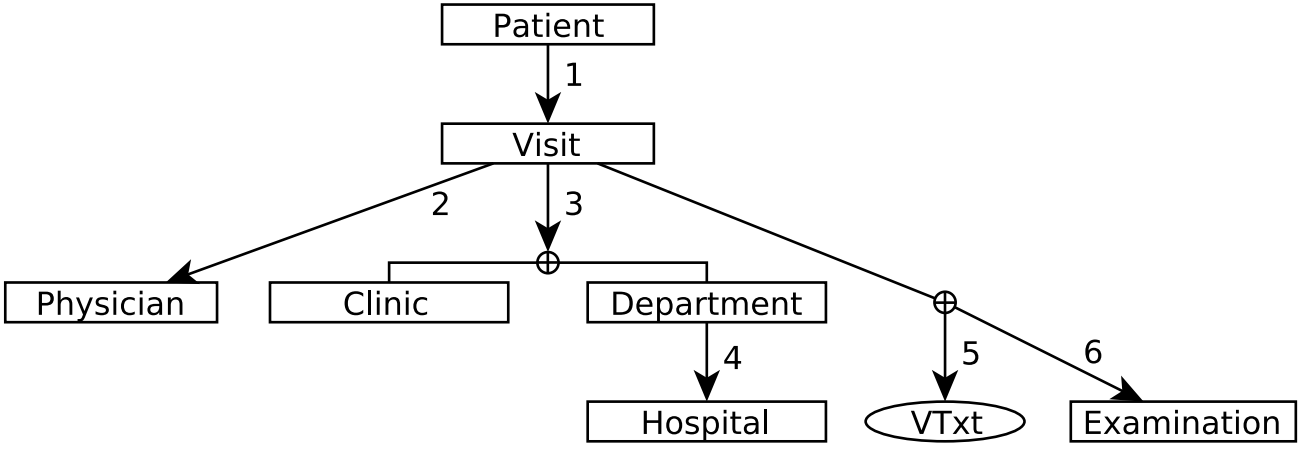
- ordering of incoming cluster types

$$order((Visit, VTxt + Examination))$$
$$order((Diagnose, DTxt + OnBase))$$

XSEM-H

- specification of a hierarchical organization of a part of a given XSEM-ER schema using hierarchical projections
- does not add any semantics
- XSEM-H schema
  - derived from an XSEM-ER schema by transformation operators
  - oriented graph
  - nodes represent entity types, relationship types, and data node types from the XSEM-ER schema
  - edges represent hierarchical projections of weak entity types and relationship types from the XSEM-ER schema and references

XSEM-H
Example



$1 : Visit[Patient \rightarrow Visit]$

$2 : Visit^{Patient}[Visit \rightarrow Physician]$

$3 : Visit^{Patient}[Visit \rightarrow Department + Clinic]$

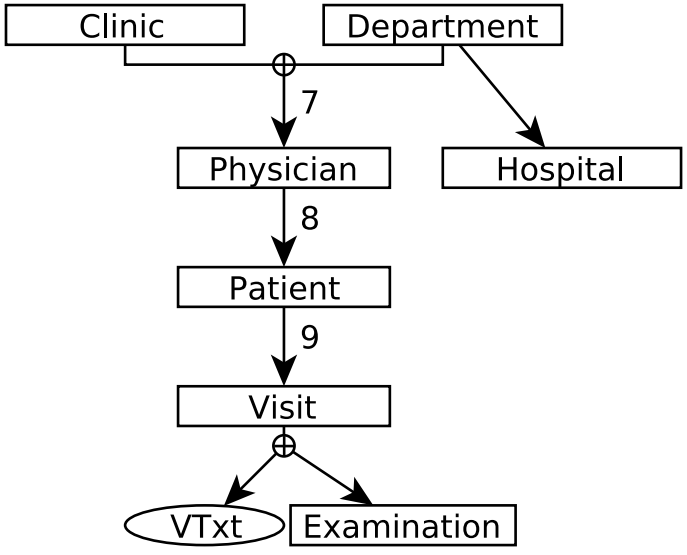
$4 : Department[Department \rightarrow Hospital]$

$5 : VTxt[Visit \rightarrow VTxt]$

$6 : Examination[Visit \rightarrow Examination]$

XSEM-H

Example



$$7 : Visit[Department + Clinic \rightarrow Physician]$$

$$8 : Visit^{Department+Clinic}[Physician \rightarrow Patient]$$

$$9 : Visit^{Department+Clinic\ Physician}[Patient \rightarrow Visit]$$

Future Work

- translation to the XML schema level
  - grammar-based languages (XML Schema, Relax NG) for describing structure
  - pattern-based languages (Schematron, XSLT, XQuery) for describing more complex integrity constraints
- translation to the logical database level
  - (object-)relational schemes
  - xml schemes
  - hybrid schemes
- extending integrity constraints
  - relativity of keys and inclusion constraints
- case tool

Publications

- Necasky M.: Conceptual Modeling for XML: A Survey. Technical Report No. 2006-3, Dep. of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, 2006, 54 p.
- Necasky M.: XSEM - A Conceptual Model for XML Data. Databases and Information Systems. Seventh International Baltic Conference on Databases and Information Systems. Communications, Materials of Doctoral Consortium. O. Vasilecas, J. Eder, A. Caplinskas (Eds.). Vilnius: Technika, July 2006, s. 328-331. ISBN 9955-28-013-1.
- Necasky M.: XSEM - A Conceptual Model for XML Data. Accepted for APCCM 2007, Ballarat, Victoria, Australia, January 2007, 12 p.