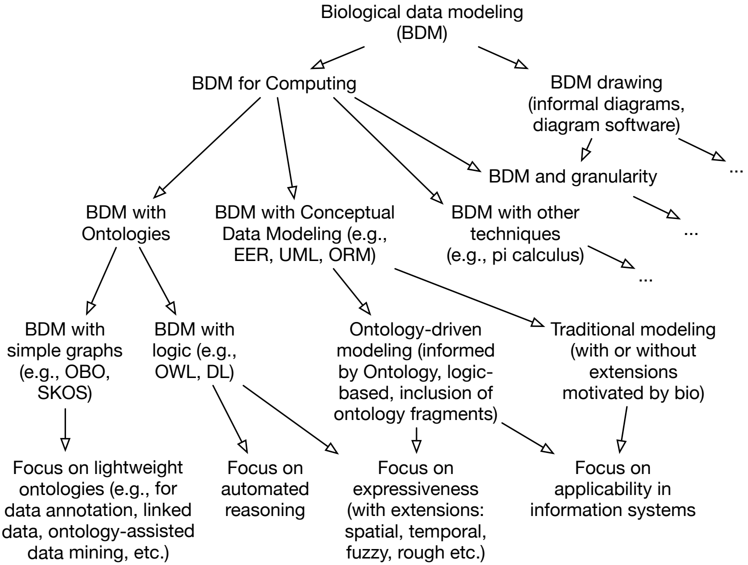
Ontology-driven formal conceptual data modelling for biological data analysis

C. Maria Keet

School of Computer Science, University of KwaZulu-Natal, South Africa
keet@ukzn.ac.za

Masters Artificial Intelligence spring School 2011 (MAIS'11)
University of KwaZulu-Natal, 26-30 September 2011

Talk based on the forthcoming book chapter in *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data*, Mourad Elloumi and Albert Y. Zomaya (Eds.). Wiley.


Biological data modeling (BDM)
BDM for Computing
BDM drawing (informal diagrams, diagram software)
BDM and granularity
...
BDM with Ontologies
BDM with Conceptual Data Modeling (e.g., EER, UML, ORM)
BDM with other techniques (e.g., pi calculus)
...
...
BDM with simple graphs (e.g., OBO, SKOS)
BDM with logic (e.g., OWL, DL)
Ontology-driven modeling (informed by Ontology, logic-based, inclusion of ontology fragments)
Traditional modeling (with or without extensions motivated by bio)
Focus on lightweight ontologies (e.g., for data annotation, linked data, ontology-assisted data mining, etc.)
Focus on automated reasoning
Focus on expressiveness (with extensions: spatial, temporal, fuzzy, rough etc.)
Focus on applicability in information systems

Motivation

- Scope: ontology-driven formal conceptual data modelling
- Audience (here at MAIS): I assume you know the motivations for conceptual data modelling and ontologies
  - Conceptual model for an *implementation*-independent view
  - Ontology for an *application*-independent view
- What we will look at:
  - Formal conceptual data modelling
  - Ontology-driven conceptual data modelling
- Both take a scientific approach to improving the quality of conceptual data models (hence, also the resulting applications), and facilitate use and reuse (interoperability)

Motivation

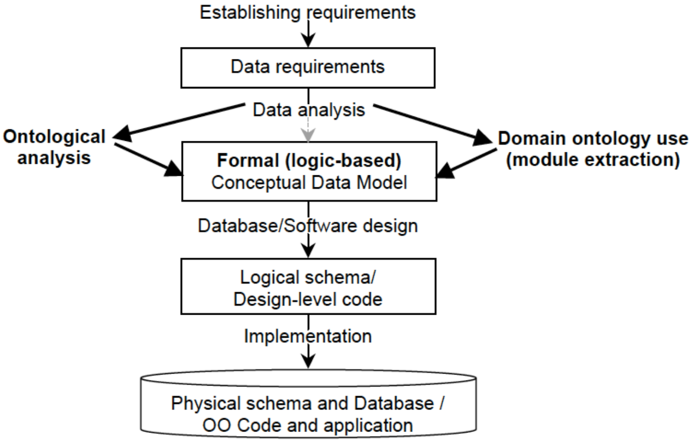
- Scope: ontology-driven formal conceptual data modelling
- Audience (here at MAIS): I assume you know the motivations for conceptual data modelling and ontologies
  - Conceptual model for an *implementation*-independent view
  - Ontology for an *application*-independent view
- What we will look at:
  - Formal conceptual data modelling
  - Ontology-driven conceptual data modelling
- Both take a scientific approach to improving the quality of conceptual data models (hence, also the resulting applications), and facilitate use and reuse (interoperability)

Motivation

- Scope: ontology-driven formal conceptual data modelling
- Audience (here at MAIS): I assume you know the motivations for conceptual data modelling and ontologies
  - Conceptual model for an *implementation*-independent view
  - Ontology for an *application*-independent view
- What we will look at:
  - Formal conceptual data modelling
  - Ontology-driven conceptual data modelling
- Both take a scientific approach to improving the quality of conceptual data models (hence, also the resulting applications), and facilitate use and reuse (interoperability)

Motivation

- Scope: ontology-driven formal conceptual data modelling
- Audience (here at MAIS): I assume you know the motivations for conceptual data modelling and ontologies
  - Conceptual model for an *implementation*-independent view
  - Ontology for an *application*-independent view
- What we will look at:
  - Formal conceptual data modelling
  - Ontology-driven conceptual data modelling
- Both take a scientific approach to improving the quality of conceptual data models (hence, also the resulting applications), and facilitate use and reuse (interoperability)


Establishing requirements
Data requirements
Data analysis
Ontological analysis
Domain ontology use (module extraction)
Formal (logic-based) Conceptual Data Model
Database/Software design
Logical schema/ Design-level code
Implementation
Physical schema and Database / OO Code and application

CDM for biological data analysis

- Many one-off bioinformatics tools (perl scripts etc.) and boutique databases
- Use of conceptual data modeling in bioinformatics limited; e.g., [BBP02, CKN$^+$10, EJF07, Kee03, PLC$^+$10, SZ05]
- Mainly EER and various type of UML diagrams
- Neither a link with ontologies (except for [EGOMA06]) nor a formal approach

Ontology-driven CDM for biological data analysis

- More precise and correct representations for correct (automated) inferences and biological knowledge discovery, surpassing human capacity (e.g., [WSH07, KRM07])
- Better data management, hence, way to make (better) use of the 'write-only' databases and 'data silos' (e.g., [CKN$^+$10])
- Reduce redundancy in scientific experiments (e.g., [MBSJ08])
- Ontological guidance for recurring modelling issues (e.g., [Kee03, AGK08, KA08, EJF07])
- Avoid adding to the pile of one-off tools
- Reusability of the information represented at the conceptual layer

1. DLs for CDM
   - $\mathcal{DLR}_{ifd}$ syntax and semantics
   - $\mathcal{CM}_{com}$
2. Extended CDM
   - Ontology-driven modeling
   - Very expressive languages
3. Automated reasoning
4. Conclusions

Introduction

Q: Which language features are 'essential' for biological data modeling?

Q: Does it make any difference which conceptual data modeling language we use for biological data modeling?

⇒ Some claim so, and different languages are used (EER, UML Class diagrams, UML Activity Diagrams, and UML Sequence Diagrams, ORM)

Q: What is the greatest common denominator (or core) of the industry-grade conceptual data modeling languages?

⇒ Compare ER, EER, UML class diagrams, ORM, and ORM2 and identify greatest common denominator. [Kee08]

- (Extends and refines [CDGL+98, CLN98, CLN99, ACK+07, Kee09])
- $\mathcal{DLR}_{ifd}$ used to formally define the generic common conceptual data modeling language $\mathcal{CM}_{com}$, i.e., with syntax and (model-theoretic) semantics, and a mapping between the two

Introduction

- Q: Which language features are 'essential' for biological data modeling?
- Q: Does it make any difference which conceptual data modeling language we use for biological data modeling?
  - ⇒ Some claim so, and different languages are used (EER, UML Class diagrams, UML Activity Diagrams, and UML Sequence Diagrams, ORM)
- Q: What is the greatest common denominator (or core) of the industry-grade conceptual data modeling languages?
  - ⇒ Compare ER, EER, UML class diagrams, ORM, and ORM2 and identify greatest common denominator: [Kee08]
  - (Extends and refines [CDGL+98, CLN98, CLN99, ACK+07, Kee09])
  - $\mathcal{DLR}_{ifd}$ used to formally define the generic common conceptual data modeling language $\mathcal{CM}_{com}$, i.e., with syntax and (model-theoretic) semantics, and a mapping between the two

Introduction

Q: Which language features are 'essential' for biological data modeling?

Q: Does it make any difference which conceptual data modeling language we use for biological data modeling?

⇒ Some claim so, and different languages are used (EER, UML Class diagrams, UML Activity Diagrams, and UML Sequence Diagrams, ORM)

Q: What is the greatest common denominator (or core) of the industry-grade conceptual data modeling languages?

⇒ Compare ER, EER, UML class diagrams, ORM, and ORM2 and identify greatest common denominator: [Kee08]

- (Extends and refines [CDGL+98, CLN98, CLN99, ACK+07, Kee09])
- $\mathcal{DLR}_{ifd}$ used to formally define the generic common conceptual data modeling language $\mathcal{CM}_{com}$, i.e., with syntax and (model-theoretic) semantics, and a mapping between the two

The base language: $\mathcal{DLR}$

Take atomic relations (**P**) and atomic concepts A as the basic elements of $\mathcal{DLR}$, which allows us to construct arbitrary relations (arity ≥ 2) and arbitrary concepts according to the syntax:

$\mathbf{R} \longrightarrow \top_n | \; \mathbf{P} \; | \; (\$i/n : C) \; | \; \neg \mathbf{R} \; | \; \mathbf{R}_1 \sqcap \mathbf{R}_2$

$C \longrightarrow \top_1 | \; A \; | \; \neg C \; | \; C_1 \sqcap C_2 \; | \; \exists[\$i]\mathbf{R} \; | \; \leq k[\$i]\mathbf{R}$

i denotes a component of a relation; if components are not named, then integer numbers between 1 and n_{max} are used, where n is the arity of the relation. Only relations of the same arity can be combined to form expressions of type $\mathbf{R}_1 \sqcap \mathbf{R}_2$, and $i \leq n$

The base language: $\mathcal{DLR}$

The model-theoretic semantics of $\mathcal{DLR}$ is specified through the usual notion of interpretation, where $\mathcal{I}= (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, and the interpretation function $\cdot^{\mathcal{I}}$ assigns to each concept C a subset $C^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ and to each n-ary $\mathbf{R}$ a subset $\mathbf{R}^{\mathcal{I}}$ of $(\Delta^{\mathcal{I}})^n$, such that the conditions are satisfied following:

$$\begin{array}{ll}
\top_n^{\mathcal{I}} \subseteq (\Delta^{\mathcal{I}})^n & (\mathbf{R}_1 \sqcap \mathbf{R}_2)^{\mathcal{I}} = \mathbf{R}_1^{\mathcal{I}} \cap \mathbf{R}_2^{\mathcal{I}} \\
\mathbf{P}^{\mathcal{I}} \subseteq \top_n^{\mathcal{I}} & (\neg C)^{\mathcal{I}} = \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}} \\
(\neg \mathbf{R})^{\mathcal{I}} = \top_n^{\mathcal{I}} \setminus \mathbf{R}^{\mathcal{I}} & (C_1 \sqcap C_2)^{\mathcal{I}} = C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}} \\
A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} & (\$i/n : C)^{\mathcal{I}} = \{(d_1, ..., d_n) \in \top_n^{\mathcal{I}} | d_i \in C^{\mathcal{I}}\} \\
\top_1^{\mathcal{I}} = \Delta^{\mathcal{I}} & (\exists[\$i]\mathbf{R})^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} | \exists (d_1, ..., d_n) \in \mathbf{R}^{\mathcal{I}}.d_i = d\} \\
 & (\leq k[\$i]\mathbf{R})^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} || \{(d_1, ..., d_n) \in \mathbf{R}_1^{\mathcal{I}} | d_i = d|\} \leq k\}
\end{array}$$

The base language: $\mathcal{DLR}$

A knowledge base is a finite set $\mathcal{KB}$ of $\mathcal{DLR}$ (or $\mathcal{DLR}_{ifd}$) axioms of the form $C_1 \sqsubseteq C_2$ and $R_1 \sqsubseteq R_2$.

An interpretation $\mathcal{I}$ satisfies $C_1 \sqsubseteq C_2$ ($R_1 \sqsubseteq R_2$) if and only if the interpretation of C_1 (R_1) is included in the interpretation of C_2 (R_2), i.e. $C_1^{\mathcal{I}(t)} \subseteq C_2^{\mathcal{I}(t)}$ ($R_1^{\mathcal{I}(t)} \subseteq R_2^{\mathcal{I}(t)}$).

$\top_1$ denotes the interpretation domain, $\top_n$ for $n \geq 1$ denotes a subset of the n-cartesian product of the domain, which covers all introduced n-ary relations.

($\$i/n : C$) denotes all tuples in $\top_n$ that have an instance of C as their i-th component.

$\mathcal{DLR}_{ifd}$

- $\mathcal{DLR}_{ifd}$ has two additional constructs compared to $\mathcal{DLR}$:
  - identification assertions on a concept C, which has the form $(\textbf{id}\ C[i_1]R_1, ..., [i_h]R_h)$, where each R_j is a relation and each i_j denotes one component of R_j.
  - Non-unary functional dependency assertions on a relation R, which has the form $(\textbf{fd}\ R\ i_1, ..., i_h \rightarrow j)$, where $h \geq 2$, and $i_1, ..., i_h, j$ denote components of R
- Syntax and semantics as for $\mathcal{DLR}$

$\mathcal{CM}_{com}$ syntax

Definition 1 (Conceptual Data Model $\mathcal{CM}_{com}$ syntax)

A $\mathcal{CM}_{com}$ conceptual data model is a tuple $\Sigma = (\mathcal{L}, \text{REL}, \text{ATT}, \text{CARD}_R, \text{CARD}_A, \text{ISA}_C, \text{ISA}_R, \text{ISA}_U, \text{DISJ}_C, \text{COVER}_C, \text{DISJ}_R, \text{KEY}, \text{EXTK}, \text{FD}, \text{OBJ}, \text{REX}, \text{RDM})$ such that:

- $\mathcal{L}$ is a finite alphabet partitioned into the sets: $\mathcal{C}$ (*class* symbols), $\mathcal{A}$ (*attribute* symbols), $\mathcal{R}$ (*relationship* symbols), $\mathcal{U}$ (*role* symbols), and $\mathcal{D}$ (*domain* symbols); the tuple $(\mathcal{C}, \mathcal{A}, \mathcal{R}, \mathcal{U}, \mathcal{D})$ is the *signature* of the conceptual data model Σ.
- REL is a function that maps a relationship symbol in $\mathcal{R}$ to an $\mathcal{U}$-labeled tuple over $\mathcal{C}$, $\text{REL}(R) = \langle U_1 : C_1, \ldots, U_k : C_k \rangle$, and k is the *arity* of R.
- ...

$\mathcal{CM}_{com}$ semantics

Definition 2 ($\mathcal{CM}_{com}$ Semantics)

Let Σ be a $\mathcal{CM}_{com}$ conceptual data model. An *interpretation* for the conceptual model Σ is a tuple $\mathcal{B} = (\Delta^{\mathcal{B}} \cup \Delta^{\mathcal{B}}_D, \cdot^{\mathcal{B}})$, such that:

- $\Delta^{\mathcal{B}}$ is a nonempty set of abstract objects disjoint from $\Delta^{\mathcal{B}}_D$;
- $\Delta^{\mathcal{B}}_D = \bigcup_{D_i \in \mathcal{D}} \Delta^{\mathcal{B}}_{D_i}$ is the set of basic domain values used in Σ; and
- $\cdot^{\mathcal{B}}$ is a function that maps:
  - Every basic domain symbol $D \in \mathcal{D}$ into a set $D^{\mathcal{B}} = \Delta^{\mathcal{B}}_{D_i}$.
  - ...
  - Every attribute $A \in \mathcal{A}$ to a set $A^{\mathcal{B}} \subseteq \Delta^{\mathcal{B}} \times \Delta^{\mathcal{B}}_D$, such that, for each $C \in \mathcal{C}$, if $\text{ATT}(C) = \langle A_1 : D_1, \ldots, A_h : D_h \rangle$, then, $o \in C^{\mathcal{B}} \rightarrow (\forall i \in \{1, \ldots, h\}, \exists a_i.$ $\langle o, a_i \rangle \in A_i^{\mathcal{B}} \wedge \forall a_i . \langle o, a_i \rangle \in A_i^{\mathcal{B}} \rightarrow a_i \in \Delta^{\mathcal{B}}_{D_i})$.
- ...

$\mathcal{CM}_{com}$ semantics

Definition 3 (Mapping $\mathcal{CM}_{com}$ into $\mathcal{DLR}_{ifd}$)

Let $\Sigma = (\mathcal{L}, \mathrm{REL}, \mathrm{ATT}, \mathrm{CARD}_R, \mathrm{CARD}_A, \mathrm{ISA}_C, \mathrm{ISA}_R, \mathrm{ISA}_U, \mathrm{DISJ}_C, \mathrm{COVER}_C, \mathrm{DISJ}_R, \mathrm{KEY}, \mathrm{EXTK}, \mathrm{FD}, \mathrm{OBJ}, \mathrm{REX}, \mathrm{RDM})$ be a $\mathcal{CM}_{com}$ conceptual data model. The $\mathcal{DLR}_{ifd}$ knowledge base, $\mathcal{K}$, mapping Σ is as follows.

- For each $A \in \mathcal{A}$, then, $A \sqsubseteq \mathtt{From}:\top \sqcap \mathtt{To}:\top \in \mathcal{K}$;
- If $C_1 \ \mathrm{ISA}_C \ C_2 \in \Sigma$, then, $C_1 \sqsubseteq C_2 \in \mathcal{K}$;
- If $R_1 \ \mathrm{ISA}_R \ R_2 \in \Sigma$, then, $R_1 \sqsubseteq R_2 \in \mathcal{K}$;
- If $U_1 \ \mathrm{ISA}_U \ U_2 \in \Sigma$, then $\mathcal{K}$ contains: $[U_1]R_1 \sqsubseteq [U_2]R_2$; $R_1 \sqsubseteq \neg R_2$;
- If $\mathrm{REL}(R) = \{U_1 : C_1, \ldots, U_k : C_k\} \in \Sigma$, then $R \sqsubseteq U_1 : C_1 \sqcap \ldots \sqcap U_k : C_k \in \mathcal{K}$;
- ...

Example: mapping to icons

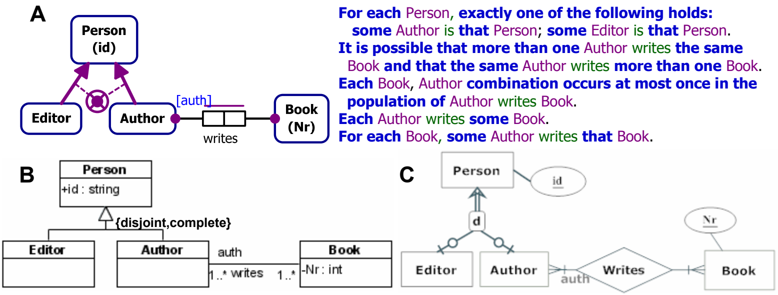


Figure: Examples of graphical syntax for $\mathcal{CM}_{com}$with ORM2 drawn in NORMA (A), UML class diagram drawn in VP-UML (B), and EER drawn with SmartDraw (C).

Example: mapping to icons

- Author ISA Person *(directed arrow in UML, EER, ORM2)*
  CARD(Author, Writes, auth) = $(1, n)$
  (1.. in UML, craw's feet and line in EER, blob and line in ORM2)*
  KEY(Person) = id *(underlined* id *in EER,* (id) *in ORM2)*
  {Author, Editor} DISJ Person
  ({disjoint} *in UML, encircled* d *in EER, encircled* X *in ORM2)*
  {Author, Editor} COVER Person
  ({complete} *in UML, open shaft arrow in EER, encircled blob in ORM2)*
- Equivalent representation in $\mathcal{DLR}_{ifd}$ as: Author $\sqsubseteq$ Person (subsumption), Author $\sqsubseteq$ $\exists$[auth]writes (at least one), Author $\sqsubseteq$ $\neg$Editor (disjoint), Person $\sqsubseteq$ Author $\sqcup$ Editor (covering), and Person $\sqsubseteq \exists^{=1}$[From]id, $\top \sqsubseteq \exists^{\leq 1}$[To](id $\sqcap$ [From] : Person) (key)

Example: mapping to icons

- $\texttt{Author}\ \textsc{isa}\ \texttt{Person}$ *(directed arrow in UML, EER, ORM2)*
  $\textsc{card}(\texttt{Author}, \texttt{Writes}, \texttt{auth}) = (1, n)$
  (1.. in UML, craw's feet and line in EER, blob and line in ORM2)*
  $\textsc{key}(\texttt{Person}) = \texttt{id}$ *(underlined* id *in EER,* (id) *in ORM2)*
  $\{\texttt{Author}, \texttt{Editor}\}\ \textsc{disj}\ \texttt{Person}$
  ({disjoint} in UML, encircled d *in EER, encircled* X *in ORM2)*
  $\{\texttt{Author}, \texttt{Editor}\}\ \textsc{cover}\ \texttt{Person}$
  ({complete} in UML, open shaft arrow in EER, encircled blob in ORM2)
- Equivalent representation in $\mathcal{DLR}_{ifd}$ as: $\texttt{Author} \sqsubseteq \texttt{Person}$ (subsumption), $\texttt{Author} \sqsubseteq \exists[\texttt{auth}]\texttt{writes}$ (at least one), $\texttt{Author} \sqsubseteq \neg\texttt{Editor}$ (disjoint), $\texttt{Person} \sqsubseteq \texttt{Author} \sqcup \texttt{Editor}$ (covering), and $\texttt{Person} \sqsubseteq \exists^{=1}[\texttt{From}]\texttt{id}$, $\top \sqsubseteq \exists^{\leq 1}[\texttt{To}](\texttt{id} \sqcap [\texttt{From}] : \texttt{Person})$ (key)

EER, UML, and ORM in terms of $\mathcal{CM}_{com}$

Definition 4

($\mathcal{CM}_{EER}$) A $\mathcal{CM}_{EER}$ conceptual data model is a tuple $\Sigma = (\mathcal{L}, \text{REL}, \text{ATT}, \text{CARD}_R, \text{ISA}_C, \text{DISJ}_C, \text{COVER}_C, \text{KEY}, \text{EXTK})$ adhering to $\mathcal{CM}_{com}$ syntax and semantics.

Definition 5

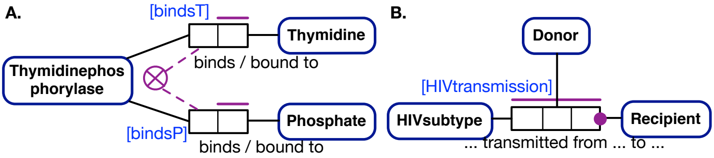
($\mathcal{CM}_{UML}$) A $\mathcal{CM}_{UML}$ conceptual data model is a tuple $\Sigma = (\mathcal{L}, \text{REL}, \text{ATT}, \text{CARD}_R, \text{ISA}_C, \text{ISA}_R, \text{DISJ}_C, \text{COVER}_C, \text{KEY}, \text{EXTK}, \text{FD}, \text{OBJ}, \text{PW})$
adhering to $\mathcal{CM}_{com}$ syntax and semantics, except for the aggregation association PW, with syntax $\text{PW} = \{U_1 : C_1, U_2 : C_2\}$, that has no defined semantics.

EER, UML, and ORM in terms of $\mathcal{CM}_{com}$

Definition 6

A $\mathcal{CM}_{ORM2^-}$ conceptual data model is a tuple
$\Sigma = (\mathcal{L}, \text{REL}, \text{ATT}, \text{CARD}_R, \text{CARD}_A, \text{ISA}_C, \text{ISA}_R, \text{ISA}_U, \text{DISJ}_C, \text{COVER}_C, \text{KEY}, \text{EXTK}, \text{FD}, \text{OBJ}, \text{DISJ}_R, \text{REX}, \text{RDM})$
adhering to $\mathcal{CM}_{com}$ syntax and semantics.

Example: constraints among relations



in $\mathcal{CM}_{com}$:

$\{\texttt{bindsT}, \texttt{bindsP}\}$ REX $\texttt{binds}$

In $\mathcal{DLR}_{ifd}$:

$\texttt{ThymidinePhosphorylase} \sqsubseteq (\exists^{\leq 1}[\texttt{bindsT}]\texttt{binds}_1 \sqcup \exists^{\leq 1}[\texttt{bindsP}]\texttt{binds}_2)$,

$[\texttt{bindsT}]\texttt{binds}_1 \sqsubseteq \neg[\texttt{bindsP}]\texttt{binds}_2$

Three main scenarios

- Provide a solution to a recurring modeling problem, informed by ontology and foundational ontologies (e.g., part-whole relations [KA08, AGK08])
- Use an ontology to generate several conceptual data models (e.g., [EGOMA06, JDM03, SS06])
- Integrate (a section of) an ontology into the conceptual data model that subsequently is converted into data in the database (e.g., KEGG, GO for annotation)

Solution to a recurring modeling problem

- (Re-)Usable components of foundational ontologies (e.g., BFO, DOLCE, GFO, ...)
  - High-level categories
  - Generic relationships (parthood, participation, dependency, constitution, etc.)
- Modeling guidance; e.g., OntoClean, ONTOPARTS
- Informs and refines language features; e.g., relational properties, role/relation-components (positionalism), keys/identification

⇒ Tells how, has justification why

Example: catalysis

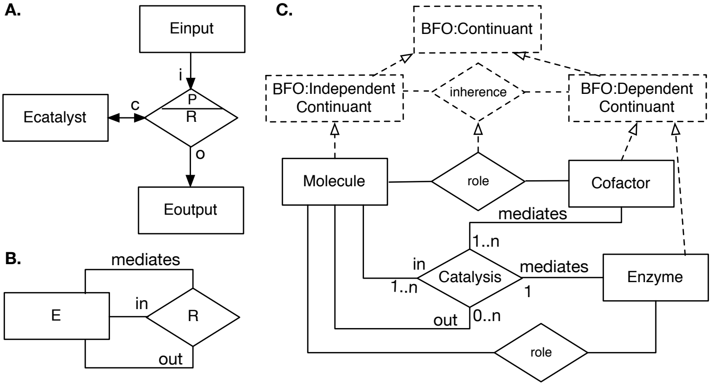
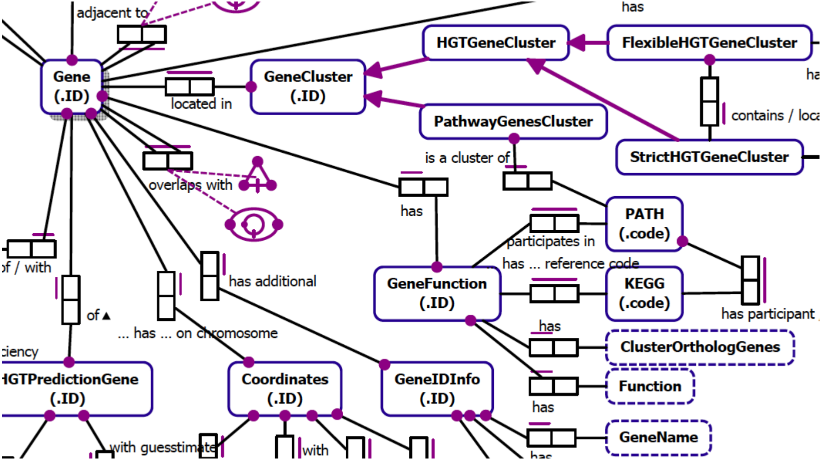
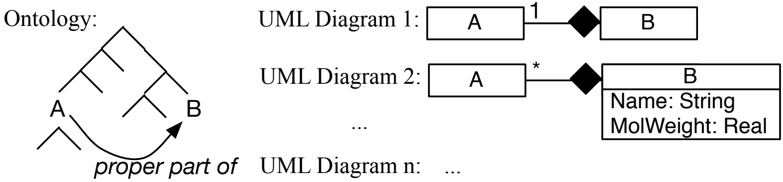


Figure: Static aspects of modeling single processes (catalytic reactions) in EER. A: Elmasri *et al.*'s [EJF07] proposal, with input, output and catalyst molecules; B: The essential roles played; C: An example of a more refined representation of catalysis,

Example: use of scenario 1 and 3



Generation of CDMs from one ontology



Need for language extensions

- Metabolic pathways (temporal)
- Central Dogma, viral infection (temporal)
- Development and transformations (temporal)
- `SmallMolecule` etc (fuzzy)
- 'typical' and default cases (probabilistic)

Temporal: examples

- Use, e.g., $\mathcal{DLR}_{\mathcal{US}}$ and $\mathcal{ER}_{\mathcal{VT}}$, with refinements on relation migration [APS07, KA10]
- $\texttt{hasRole}$ $\textsc{RDex}$ $\texttt{autocatalysis}$

  $\texttt{R}$ $\textsc{RDex}$ $\texttt{R}'$ if and only if $\langle o_1, o_2\rangle \in \texttt{R}^{\mathcal{I}(t)} \rightarrow \exists t' > t.\langle o_1, o_2\rangle \in \texttt{R}'^{\mathcal{I}(t')}$

  $\textsc{rel}(\texttt{hasRole}) = \{\texttt{bearer} : \texttt{RNAmolecule}, \texttt{role} : \texttt{Ribozyme}\}$

  $\textsc{rel}(\texttt{autocatalysis}) = \{\texttt{substrate} : \texttt{RNAmolecule}, \texttt{catalyst} : \texttt{Ribozyme}\}$
- $\texttt{Monocyte}$ $\textsc{dev}$ $\texttt{Macrophage}$

  $\texttt{Monocyte} \sqsubseteq \Diamond^{+}(\texttt{Macrophage} \sqcap \neg\texttt{Monocyte})$
- Viral entry: binding, membrane fusion or detach, viral entry (details in chapter)

Standard reasoning services

- CDM consistency
- Class consistency
- Class subsumption
- Refinement of multiplicities
- (instance classification and retrieval)
- Examples: discovery of new protein phosphatase, quickly finding suitable rubber or pharma molecules

Standard reasoning services

- CDM consistency
- Class consistency
- Class subsumption
- Refinement of multiplicities
- (instance classification and retrieval)
- Examples: discovery of new protein phosphatase, quickly finding suitable rubber or pharma molecules

Standard reasoning services

- CDM consistency
- Class consistency
- Class subsumption
- Refinement of multiplicities
- (instance classification and retrieval)
- Examples: discovery of new protein phosphatase, quickly finding suitable rubber or pharma molecules

Querying and other reasoning services

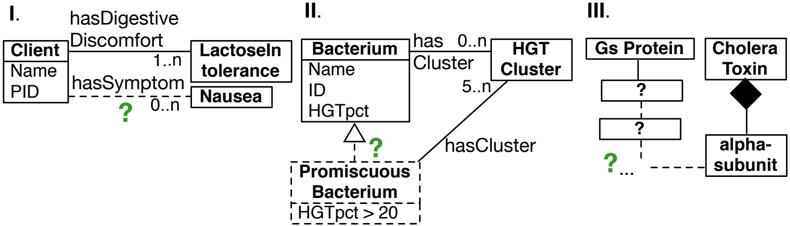


Figure: Graphical depictions of the three query patterns to find 'new' classes or relationships supported by the data; (i): correlation; (ii): hypothesis about existence of subclass PromiscuousBacterium; (iii): path query to check whether the Gs protein somehow relates to the alpha-subunit of the CholeraToxin.

Conclusions and future directions

- Substantiated advantages of ontology-driven formal conceptual data modeling:
  - Formal foundation for UML, EER, ORM, with $\mathcal{CM}_{com}$, which has an equi-satisfiable $\mathcal{DLR}_{ifd}$ knowledge base
  - Ontological guidance to motivate better modeling choices, illustrated with a refinement for representing catalytic reactions
  - Claimed to be 'non-representable' biological knowledge can be represented in $\mathcal{CM}_{com}$ (n-aries, constraints among relationships)
- Language extensions for, a.o., temporal knowledge; demonstrated with related processes in a cascade of reactions
- Automated reasoning services were illustrated for taxonomic classification
- Three different query patterns to find new type-level information

Conclusions and future directions

- Substantiated advantages of ontology-driven formal conceptual data modeling:
  - Formal foundation for UML, EER, ORM, with $\mathcal{CM}_{com}$, which has an equi-satisfiable $\mathcal{DLR}_{ifd}$ knowledge base
  - Ontological guidance to motivate better modeling choices, illustrated with a refinement for representing catalytic reactions
  - Claimed to be 'non-representable' biological knowledge can be represented in $\mathcal{CM}_{com}$ (n-aries, constraints among relationships)
- Language extensions for, a.o., temporal knowledge; demonstrated with related processes in a cascade of reactions
- Automated reasoning services were illustrated for taxonomic classification
- Three different query patterns to find new type-level information

Conclusions and future directions

- Substantiated advantages of ontology-driven formal conceptual data modeling:
  - Formal foundation for UML, EER, ORM, with $\mathcal{CM}_{com}$, which has an equi-satisfiable $\mathcal{DLR}_{ifd}$ knowledge base
  - Ontological guidance to motivate better modeling choices, illustrated with a refinement for representing catalytic reactions
  - Claimed to be 'non-representable' biological knowledge can be represented in $\mathcal{CM}_{com}$ (n-aries, constraints among relationships)
- Language extensions for, a.o., temporal knowledge; demonstrated with related processes in a cascade of reactions
- Automated reasoning services were illustrated for taxonomic classification
- Three different query patterns to find new type-level information

Outlook

- Ontology-driven formal conceptual data modeling is still a relatively young field
- How to handle incomplete information in hypothesis testing [Kee10]?
- how OntoClean [GW04] ideas can be incorporated in conceptual data modeling methodologies
- Formal link between ontologies and conceptual data models
- Development of CASE tools with both a unifying formalism and an integrated automated reasoner, and multiple language interfaces
- Temporal reasoning beyond $\mathcal{ER}_{\mathcal{VT}}$ and its $\mathcal{DLR}_{\mathcal{US}}$ foundation, principally either as extension to UML Class Diagrams or as formalization of Sequence and Activity Diagrams

References

Alessandro Artale, Diego Calvanese, Roman Kontchakov, Vladislav Ryzhikov, and Michael Zakharyaschev.
Reasoning over extended ER models.
In *ER-07*, volume 4801 of *LNCS*, pages 277–292. Springer, 2007.

Alessandro Artale, Nicola Guarino, and C. Maria Keet.
Formalising temporal constraints on part-whole relations.
In Gerhard Brewka and Jerome Lang, editors, *11th International Conference on Principles of Knowledge Representation and Reasoning (KR'08)*, pages 673–683. AAAI Press, 2008.
Sydney, Australia, September 16-19, 2008.

A. Artale, C. Parent, and S. Spaccapietra.
Evolving objects in temporal information systems.
Annals of Mathematics and Artificial Intelligence, 50(1-2):5–38, 2007.

E. Bornberg-Bauer and N.W. Paton.
Conceptual data modelling for bioinformatics.
Briefings in Bioinformatics, 3(2):166180, 2002.

D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati.
Description logic framework for information integration.
In *Proc. of the 6th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR98)*, pages 2–13, 1998.

D. Calvanese, C. M. Keet, W. Nutt, M. Rodríguez-Muro, and G. Stefanoni.
Web-based graphical querying of databases through an ontology: the WONDER system.
In *Proceedings of ACM Symposium on Applied Computing (ACM SAC'10)*, pages 1389–1396. ACM, 2010.
March 22-26 2010, Sierre, Switzerland.

References

D. Calvanese, M. Lenzerini, and D. Nardi.
Logics for Databases and Information Systems, chapter Description logics for conceptual data modeling.
Kluwer, Amsterdam, 1998.

D. Calvanese, M. Lenzerini, and D. Nardi.
Unifying class-based representation formalisms.
Journal of Artificial Intelligence Research, 11:199–240, 1999.

H. El-Ghalayini, M. Odeh, R. McClatchey, and D. Arnold.
Deriving conceptual data models from domain ontologies for bioinformatics.
In *2nd Conference on Information and Communication Technologies (ICTTA'06)*, pages 3562 – 3567. IEEE Computer Society, 2006.

Ramez Elmasri, Feng Ji, and Jack Fu.
Modeling biomedical data.
In Jake Chen and Editors Amandeep S. Sidhu, editors, *Biological database modeling*, chapter 3. Artech House Publishers, 2007.

Nicola Guarino and Chris Welty.
An overview of OntoClean.
In S. Staab and R. Studer, editors, *Handbook on ontologies*, pages 151–159. Springer Verlag, 2004.

M. Jarrar, J. Demy, and R. Meersman.
On using conceptual data modeling for ontology engineering.
Journal on Data Semantics: Special issue on Best papers from the ER/ODBASE/COOPIS 2002 Conferences, 1(1):185–207, 2003.

References

C. Maria Keet and Alessandro Artale.
Representing and reasoning over a taxonomy of part-whole relations.
Applied Ontology – Special issue on Ontological Foundations for Conceptual Modeling, 3(1-2):91–110, 2008.

C. Maria Keet and Alessandro Artale.
A basic characterization of relation migration.
In R. Meersman et al., editors, *OTM Workshops, 6th International Workshop on Fact-Oriented Modeling (ORM'10)*, volume 6428 of *LNCS*, pages 484–493. Springer, 2010.
October 27-29, 2010, Hersonissou, Crete, Greece.

C. Maria Keet.
Biological data and conceptual modelling methods.
Journal of Conceptual Modeling, 29, October 2003.
http://www.inconcept.com/jcm.

C. Maria Keet.
A formal comparison of conceptual data modeling languages.
In *13th International Workshop on Exploring Modeling Methods in Systems Analysis and Design (EMMSAD'08)*, volume 337 of *CEUR-WS*, pages 25–39, 2008.
16-17 June 2008, Montpellier, France.

C. Maria Keet.
Mapping the Object-Role Modeling language ORM2 into Description Logic language $\mathcal{DLR}_{ifd}$.
Technical Report arXiv:cs.LO/0702089v2, KRDB Research Centre, Free University of Bozen-Bolzano, Italy, April 2009. arXiv:cs.LO/0702089v2.

References

C. Maria Keet.
Ontology engineering with rough concepts and instances.
In P. Cimiano and H.S. Pinto, editors, *17th International Conference on Knowledge Engineering and Knowledge Management (EKAW'10)*, volume 6317 of *LNCS*, pages 507–517. Springer, 2010.
11-15 October 2010, Lisbon, Portugal.

C. Maria Keet, Marco Roos, and M. Scott Marshall.
A survey of requirements for automated reasoning services for bio-ontologies in OWL.
In *Proceedings of the 3rd Workshop on OWL: Experiences and Directions (OWLED 2007)*, volume 258 of *CEUR-WS*, 2007.
6-7 June 2007, Innsbruck, Austria.

Joshua S. Madin, Shawn Bowers, Mark P. Schildhauer, and Matthew B. Jones.
Advancing ecological research with ontologies.
Trends in Ecology & Evolution, 23(3):159–168, 2008.

Oscar Pastor, Ana M. Levin, Juan Carlos Casamayor, Matilde Celma, Luis E. Eraso, Maria Jose Villanueva, and Manuel Perez-Alonso.
Enforcing conceptual modeling to improve the understanding of human genome.
In *Fourth International Conference on Research Challenges in Information Science (RCIS'10)*, pages 85–92. IEEE Computer Society, 2010. Nice, France, 19-21 May 2010.

References

Vijayan Sugumaran and Veda C. Storey.
The role of domain ontologies in database design: An ontology management and conceptual modeling environment.
ACM Transactions on Database Systems, 31(3):1064–1094, 2006.

Daniel Shegogue and W. Jim Zheng.
Object-oriented biological system integration: a SARS coronavirus example.
Bio, 21(10):25022509, 2005.

K. Wolstencroft, R. Stevens, and V. Haarslev.
Applying OWL reasoning to genomic data.
In C.J.O. Baker and H. Cheung, editors, *Semantic Web: revolutionizing knowledge discovery in the life sciences*, pages 225–248. Springer: New York, 2007.

Thank you for your attention