Knowledge and Data Integration
2017-2018

Department of Information Engineering and Computer Science

Modeling Conceptual Models – Step 1

Document data:	**Reference persons:**
16.10.2017	Tovo Alessia 189192; Luca Massimiliano 173497; Tavonatti Stefano 183846

forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.

Index

Revision History

Revision	Date	Author	Description of Changes
0.1	16.10.2017	Luca Massimiliano	Document Created and layout section deleting
0.2	19.10.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano	Sections review
0.3	23.10.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano	Assignment description
0.4	24.10.2017	Tovo Alessia	Personas review
0.5	24.10.2017	Tavonatti Stefano	Scenario review
0.6	24.10.2017	Luca Massimiliano	Storytelling review
0.7	24.10.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano	General review
0.8	25.10.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano	Query and EER diagram
1.0	25.10.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano; Mattia Fumagalli; Fausto Giunchiglia	Meeting and general review
1.1	28.10.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano	New query and Persona after the review
1.2	29.10.2017	Tavonatti Stefano	Persona added
1.3	29.10.2017	Tovo Alessia	Persona added and scenario review

1.4	29.10.2017	Luca Massimiliano	Persona added
1.5	30.10.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano	Meeting and general review
1.6	31.10.2017	Luca Massimiliano	Add references and description review
1.7	03.11.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano	Meeting and EER diagram review
1.8	03.11.2017	Luca Massimiliano	Query specification
1.9	03.11.2017	Tovo Alessia	Query review
1.10	04.11.2017	Tavonatti Stefano	Query review
1.11	06.11.2017	Luca Massimiliano; Tovo Alessia; Tavonatti Stefano	EER description and general review

1. Assignment description

In this project, our team would like to provide a brand new solution to allow people to find out the best semi-formal way to model something. Our idea is to provide a top-level ontology and an open source search engine that allow people to look up for the best conceptual model to use. It is not always simple to choose the right model to use and actually there are no tools that support student, researcher or professors or whoever want to semi-formalize something and so we decided to build it. There are basically six way to model things: data-flow modeling (DFM), entity-relationship modeling (ERM), event-driven process chain (EPC), joint application development (DSDM), place/transition net and finally state-transition modeling. A big issue is that there is not a tool that allows designers to choose the best conceptual model and we also find out that in most of the cases there is more than one model that can provide a valid solution and choose between them should not be simple. In order to provide a valid and valuable solution we would like to do basically two things: focus on the versioning of ER (it has been done only on ER just because we do not have enough resources - time - to look also at other conceptual models) and show how the EER provided at the end of this document and the followed ontology can be easily used also for all the other conceptual models. For this reason, if you look at the EER you can find out that the name of all the entities is as generic as possible and precise and meaningful at the same time.

To build this tool we found out that there are not structured or semi structured datasets that can help us and so all the data has been retrieved manually or automatically (RapidMiner ©) by web pages or papers (pdf). We decided to try to map in our EER schema the ER model and to do the versioning on all the different standard of representation.

2. Scenario

At the University of Santa Barbara, there are three kind Persona that are interested in finding out which are the differences between the big variety of standard representations allowed in ER and also to find out the more useful and helpful to achieve a specific problem. All the Persona listed below really exists in their role. Doctor Bang is a master student, doctor Drawert is a PostDoc researcher in UCSB and professor Belding teach at the Computer Science Department of UCSB.

- a student, Lucas Bang
- a researcher, Brian Drawert
- a professor, Elizabeth M. Belding

The Santa Barbara University has a Computer Science Department. So Lucas Bang is a master student in Computer Science and has to prepare a project for one of his exams. This project could also become the thesis project, so the work he is doing must be as precise and refined as possible.

Since the main argument of the project is the ER model and his applications in different domains, he has to better know how ER maps domains. Especially he has to study the different versions of the ER model.

Since there is not a standard documentation of ER, Lucas needs to use a tool that allows him to find all the possible information about all the different existing versions of ER.

Brian Drawert is a researcher at the Department of Information Engineering and Computer Science and he wants to prepare a paper in which he analyzes the evolution of the ER model.

Actually there are not resources that can help him to do this kind of work and with the state of the art, he has to do it totally manually and this implies that he will take a long time. Provide a tool that enables this kind of work, will dramatically reduce the time needed to do this kind of work.

Elizabeth M. Belding is a professor and cares about the field of data integration. She is looking for a standard representation that is as close as possible to the OWL and RDF constraint in order to allow a faster and better conversion to those two.

She would like to analyze which standard enable what. For instance, the min-max ISO is a representation that allows defining the inverse relation while the Chen standard does not provide any constraint to represent it. Having a tool that allow you to look at the constraint allowed by a standard (and to compare two standards of the same conceptual model) will reduce the amount of time needed and will ensure a better result (as long as you can query things like “provide me the standard that has constraint 1 and constraint 2 and … “)

These three Personas have, at the same time, different and similar backgrounds. In fact, they all are in the Computer Science field but have different experiences and different knowledge about Conceptual Models and particularly about ER.

Moreover, they have different purposes to use this tool, but all their purposes can be mapped into the tool provided since it gives all the possible information that could interest an “academic” person.

3. Personas

- Lucas Bang - Student
- Brian Drawert - Researcher
- Elizabeth M. Belding - Professor

Lucas Bang:

Lucas Bang is a Computer Science student. He is a Master Degree student and he needs to design an ER diagram in order to semi-formalize a part of a project that has to be very well done since it could be extended to a thesis.

Lucas doesn't know very much about ER model. He knows some constructs, for which purpose they are used, which is the domain where ER model works. He would like to know more about ER since he has to build a project on it. For example, Lucas would like to know more in detail ER's versions, which are the main differences, for what they are used, if is better to use a certain version for a specific scenario, if ER could be mapped into a particular programming language.

Since there is not a standard documentation about ER, Lucas needs to use this tool in order to find more information as possible, in a simple, synthetic and fast way.

Brian Drawert:

Brian Drawert is a researcher of the Department of Computer Science of UCSB. He works in the field of Knowledge and Data integration and he works in a team which analyzes different conceptual models. He is currently working on a project in which he studied the evolution of the ER model. In particular, he focuses on the difference between the various representation of the model *and the different output*s.

Brian knows that there is not a standard representation of this model, so he needs to use this tool in order to speed up its work and find more information as possible.

Elizabeth M. Belding:

Elizabeth Belding is professor and vice chair at the Computer Science Department of the University of Santa Barbara. She is researching on data modeling, database and data integration and so at the conceptual models that can help to define schemas for those fields. Clearly it is not simple with the state of the art to choose the best version of ER that allows you to move on.

Of course she well-know the state of the art but at the same time, it should be helpful for her to have a tool that allows her to visualize the differences between two version (also in order to have a quick overview on it).

4. Storytelling definition

Persona	Real World action by persona	System/demo action
Lucas Bang	Search the differences Chen's version and Bachman's version	Return the Conceptual Model ER with construct defined in Chen's version but not in Bachman's version
Lucas Bang	Search all the common elements between IDEF1X version and Martin's version	Returns the common characteristics of the two versions
Lucas Bang	Search a complete documentation of Chen's representation of ER	Returns all the entities of the model related to Chen ER representation
Lucas Bang	Search for the original papers of Chen's representation in order to cite it in its paper	Returns the conceptual model with all the information related to the author
Brian Drawert	Search for the different documents that can be produced by the conceptual model	Returns the conceptual model with all information about the correlated documents producible
Brian Drawer	Search the conceptual models maintained by the organization W3C	Returns the conceptual models maintained by the organization W3C
Brian Drawer	Search all the conceptual models available in the Database modeling domain	Returns all the conceptual model available in the Database modeling domain
Brian Drawer	Search for all the conceptual models which as AplicationDomain = Database modeling and has the inverse relationship constraint	Returns the conceptual models in the Database modeling application domain which has the inverse relationship constraint.
Elizabeth Belding	Looking for something that is standardized so her work will be appreciated and understood by all the world	Returns all the version of ER that are standardized and maintained by someone
Elizabeth Belding	After she finds out all the standard versions, she would like to filter the versions which have the direction constraint	Returns min-max ISO and UML as long as are the only two that are standard and allow you to represent the direction of the relation

Elizabeth Belding	Look for conceptual models that have Database as domain and that the document produced is human readable	Returns the conceptual models in the Database application domain which document produced is human readable
Elizabeth Belding	Search all the Conceptual models that work in Database domain and can be used with programming language SQL	Returns all the conceptual model s that works in Database domains and can be mapped to SQL language

Tab. 1

5. Queries and generalized queries

As long as we believe that all the information stored should be useful and valuable, when we write that we will return a Conceptual Model. We will return also authors, constraints and other information about that specific CM (or a specific version of a CM). The fastest way to compute the query in this scenario is to have a View built as ConceptualModel ⋈ Version ⋈ (ApplicationDomain ⋈ Construct ⋈ DocumentProduced ⋈ Author ⋈ Representation ⋈ PLCategory) which is called ConceptualModelInformationContainer (we will refer to it with CMIC). Another way to achieve our goal is to do this join basically in all the query we designed or to return less information. In the following table we assume that the view exists.

Persona	(Generalized) Query	(Expected) Answer
Lucas Bang	Give me the differences between Version.name = x of ConceptualModel.name = y and Version.name = z of ConceptualModel.name = y	`(SELECT * FROM CMIC WHERE Version.name = x AND ConceptualModel.name = y) MINUS (SELECT * FROM CMIC WHERE Version.name = z AND ConceptualModel.name = y)`
Lucas Bang	Give me the similarities between Version.name = x of ConceptualModel.name = y and Version.name = z of ConceptualModel.name = y	`(SELECT * FROM CMIC WHERE Version.name = x AND ConceptualModel.name = y) INTERSECT (SELECT * FROM CMIC WHERE Version.name = z AND ConceptualModel.name = y)`
Lucas Bang	Give me the ConceptualModel of Version.name=x	`SELECT * FROM CMIC WHERE Version.name = x`
Lucas Bang	Give me the ConceptualModel that has as definition the paper Author.name = x	`SELECT * FROM CMIC WHERE Author.name = x`

Brian Drawert	Give me a ConceptualModel that allows me to produce DocumentProduced.name = x	`SELECT * FROM CMIC WHERE DocumentProduced.name = x`
Brian Drawert	Give me a ConceptualModel that is maintained by Organization.name = x	`SELECT * FROM CMIC WHERE Organization.name = x`
Brian Drawert	Give me all the ConceptualModel available in the ApplicationDomain.name = x	`SELECT * FROM CMIC WHERE ApplicationDomain.name = x`
Brian Drawert	Give me all the ConceptualModel that stand in ApplicationDomain.name = x and allow to use Constraint.name = y	`SELECT * FROM CMIC WHERE ApplicationDomain.name = x AND Constraint.name = y`
Elizabeth Belding	Give me all the ConceptualModel that has Version.isStandard = true	`SELECT * FROM CMIC WHERE Version.isStandard = 1`
Elizabeth Belding	Give me all the ConceptualModel that has Version.isStandard = true and Version.isMainteined = true provide Construct.name = x	`SELECT * FROM CMIC WHERE Version.isStandard = 1 AND Version.isMainteined = 1 AND Construct.name = x`
Elizabeth Belding	Give me all the ConceptualModel that stand in ApplicationDomain.name = x and DocumentProduced.isHumanReadable = true	`SELECT * FROM CMIC WHERE ApplicationDomain.name = x AND DocumentProduced.isHumanReadable = 1`
Elizabeth Belding	Give me all the ConceptualModel that stand in ApplicationDomain.name = x and PLCategory.name = y	`SELECT * FROM CMIC WHERE ApplicationDomain.name = x AND PLCategory.name = y`

Tab. 2

6. Data Sets

- https://en.wikipedia.org/wiki/Entity–relationship_model
- https://www.smartdraw.com/entity-relationship-diagram/
- https://pdfs.semanticscholar.org/3aa6/b6e378fc17b64f4920e2b5240be9a30477ca.pdf
- http://bit.csc.lsu.edu/~chen/pdf/erd-5-pages.pdf
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.4630&rep=rep1&type=pdf
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.280&rep=rep1&type=pdf
- http://tinman.cs.gsu.edu/~raj/4710/f11/Ch08.pdf
- http://web.cs.ucdavis.edu/~green/courses/ecs165a-w11/2-er.pdf
- http://www.is.informatik.uni-kiel.de/thalheim/HERM/HERMindetail.pdf
- http://www.peter-lo.com/Teaching/IT354/L02.pdf

7. Extended Entity-Relationship Model (EER)

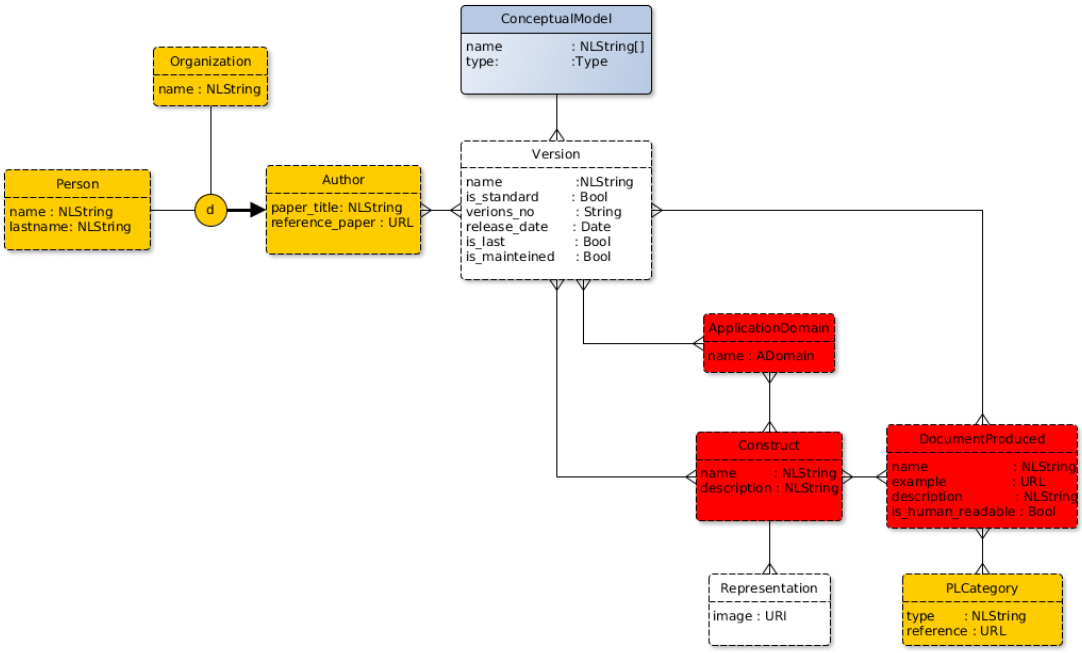

Fig. 1. The EER diagram used to model the problem

The color meaning is:

- White : complex attributes
- Red : auxiliary entities
- Blu : core entities
- Yellow : common entities

There is one core entity which is called **ConceptualModel** and it is the entity that we always return. In order to provide more meaningful information, we actually return the join over all the table but everything is based on the **ConceptualModel** table. All the other entities are used to apply the right filters to the query that will be launched by the user. **Version** has to be considered as a complex attribute of **ConceptualModel** and is crucial for our system. It contains information about the conceptual model version and it allows us to create a filter that is fundamental.

Suppose you are looking for a specific version of ER: Chen's representation allows you to represent entities, relationships and attributes while the min-max allows you to represent also the direction of the relationship. It means that, thanks to this table, you will be able to do two things: take a look on how the model has evolved and also know which version of ER model use to define create your diagram. For this reason, as you can see from the diagram all the other filter are not directly linked to **ConceptualModel** but pass through **Version**. Given the domain, we can also

describe which programming language paradigm (the **PLCategory** common entity) can be mapped on it (if there is one). For example, when the result is ER and UML is the version, the mappable programming languages are the object oriented programming languages while when the version is Chen, as long as the domain is “Database Design”, the mappable languages are all the Query Languages

Of course we provide also the **constructs** linked to a specific **Version** of a **ConceptualModel** and we also assume that the same constraint can have multiple representation as long as some conceptual model does not have a standard. Just think about ER: there couple of ways to represent entities and relationship (particularly the problem is on the representation of the cardinality).

There is also a common entity which is shared among different schemas: **Author**. We use this entity to define who takes care of the standard definition or the development of the Conceptual Model. Author is further specialized in **Organization or Person**

As said, all the information that we retrieve for building this schema are non-structured so all the decisions are based on the analysis of texts and reviews that are done in weekly meetings. Some of the decision taken for modeling the database schema are assumptions:

- If you look to something you do not specify the version, then you are referring to the Chen’s version as long as it is the most widely used.
- The same Author take care of a **ConceptualModel** from the very beginning to the end and this implies that there is only one organization linked to a conceptual model and this organization can not change in time.