**Dipartimento di Ingegneria e Scienza dell’Informazione**

KDI Lab 2018-19

|  |  |
| --- | --- |
| **Document data:** | **Reference persons:** |
| dd.mm.yyyy | Name, Familiy Name - Student ID; … |

© 2018 University of Trento

Trento, Italy

KnowDive (internal) reports are for internal only use within the KnowDive Group. They describe preliminary or instrumental work which should not be disclosed outside the group. KnowDive reports cannot be mentioned or cited by documents which are not KnowDive reports. KnowDive reports are the result of the collaborative work of members of the KnowDive group. The people whose names are in this page cannot be taken to be the authors of this report, but only the people who can better provide detailed information about its contents. Official, citable material produced by the KnowDive group may take any of the official Academic forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.

**Index**

[1. Preliminary information 4](#_Toc525546098)

[2. Data Preparation and Analysis (difficulty: “Medium”) 4](#_Toc525546099)

[3. Domain (Informal) Modelling (difficulty: “Low/Medium”)](#_Toc525546103)

[4. Domain (Formal) Modelling (difficulty: “Medium/High”) 7](#_Toc525546107)

[5. Data Integration (difficulty: “High”) 9](#_Toc525546111)

**Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Revision** | **Date** | **Author** | **Description of Changes** |
| 0 | dd.mm.yyyy | Name | Text |
| … | … | … | … |

**Layout**

* **Paragraphs**. *Times Roman* *size 12* should with *line spacing 1,0* be used.
* **Sections**. *Times Roman* *size 14.*
* **Title**. *Times Roman* *size 28.*
* **Footnotes**. Put footnotes at the bottom of the page and use 9 points text. They may be numbered or referred to by asterisks or other symbols.[[1]](#footnote-1) Footnotes should be separated from the main text by a line.[[2]](#footnote-2)
* **Illustrations.** Place figures, tables, and photo-graphs in the paper near where they are first discussed, rather than at the end, if possible.
* **Captions.** Provide a caption for every illustration; number each one sequentially in the form: “Figure 1. Caption of the Figure.” “Table 1. Caption of the Table.” Type the captions of the figures and tables below the body, using 11 point text.
* *Margins and Gaps are already defined by the current doc.*

# Preliminary information

* In what follows we provide **4 types of assignment**. Each of them is related to some steps of the methodology explained during the course.
* Each assignment refers to topics that are described all over the course, thus it is **mandatory to follow each class** during the course.
* Students are supposed to select one of the topics and then address the corresponding tasks, producing a document that is an exploitation of one the sections below and, eventually, generating the requested output files.
* During the course two **progress meetings** are scheduled before the final examination. It is up to the students to use these progress meetings to fix bugs and to improve their works.
* The **final examination** will be an assessment of the output document, the produced files and the way by which students interacted during the progress meetings.
* Notice that, for each topic, we assigned a **difficulty rank**. This will affect the final scoring (e.g., if students chose a low difficulty assignment the quality of the output has to be very high in order to get a full score).

# Data Preparation and Analysis (difficulty: “Medium”)

# *Assignment Structure*

* **Input** 
  + Topic
  + Dataset(s)
* **Tool(s)**
  + Google/Google Data Search
  + Rapid Miner
  + Custom scripts
* **Output**
  + Tech report (the structure is described below, sub-section “c”)
  + Cleaned Data (.xls file)

# *Process Description*

1. **Select a topic and search/generate related dataset(s).** Notice that dataset(s) *i)* can be searched (google and google data search are the suggested tools) or *ii)* can be generated by using RapidMiner or a custom script and by scraping information from web pages (this will be taken into account for the final scoring).
2. **Clean and/or merge and/or refine data.** For this task the usage of RapidMiner is mandatory.
3. **Analyze the data.** For this task the use of RapidMiner is mandatory. Notice that we allow the students to choose how to analyze the data by selecting some of the techniques explained during the classes.
4. **Generate the technical report.** (Guidelines below, sub-section “c”).

# *Tech Report Description*

* **Section 1 – [Scenario (and Personas) description]**.Describe a possible example of domain application by providing an example of scenario application. The scenario is a description of the situation that tells how the storytelling will be developed. It is in plain text. It is like telling a story to a kid describing a real world scenario with real needs, real people, real action. People should have names, an age, gender specified, roles, and should take action. *The scenario description should not exceed the maximum of a page and be at least ½ page*.
* **Section 2 – [Queries description]**.Starting from the personas, formalize a complete list of queries with example results (taken from the dataset).This section should not be longer than 2 pages (and not be shorter than 1 page).
* **Section 3 – [Dataset(s) extraction/generation and Dataset(s) description]**. Here a description of the process by which the data were collected/generated should be provided, by listing source links (data, custom scripts and RapidMiner.xml). *This section should not be longer than a page and not shorter than 1 page.*
* **Section 4 – [Dataset(s) cleaning, merging and analysis description]**. Here a description of the process by which data were cleaned (and merged) and of the results of the analysis (along with charts) should be provided. Moreover, links to the output dataset(s) (i.e., the result of the cleaning) should be provided. *This section should not be longer than 3 pages (and not be shorter than 2 pages).*
* **Section 5 – [Related ontology proposal]**.Considering the datasets, provide a list of (maximum 5 and minimum 3) ontologies that can be used for representing that kind of information. Make a ranking of the ontologies and motivate the choices. *This section should not be longer than 3 pages (and not be shorter than 2 pages).*
* **Section 6** **– [Related datasets proposal and integration example]**. Provide a list of datasets that can be integrated with the source input dataset and describe how it can be done. *This section should not exceed the maximum of 2 page and be at least 1 page*.
* **Section 7** – **[Final considerations and open issues]**.

# Domain (Informal) Modelling (difficulty: “Low/Medium”) con formal

# *Assignment Structure*

* **Input** 
  + Topic and/or Tech Report(s)
* **Tool(s)**
  + Google/Google Scholar
  + yED
  + MySQL WorkBench
  + Protégé (optional)
* **Output**
  + Tech report (the structure is described below, sub-section “c”)
  + yED file and/or WorkBench file

# *Process Description*

1. **Select a topic and/or tech report.** Notice that topic(s) and tech reports *i)* can be provided by the course designers or *ii)* can be provided by the students. In the second case a proposal should be submitted to the course designers for approval.
2. **Design the model.** For this task it is possible to use yEd or MySQL Workbench. Notice that we allow the students to choose what tool to use (the usage of MySQL WorkBench is considered more difficult and this will affect the final score)
3. **Generate the technical report.** (Guidelines below, sub-section “c”).

# *Tech Report Description*

* **Section 1 – [Scenario description]**. The scenario is a description of the situation that tells how the storytelling will be developed. It is in plain text. It is like telling a story to a kid describing a real world scenario with real needs, real people, real action. People should have names, an age, gender specified, roles, and should take action. *The scenario description should not exceed the maximum of a page and be at least ½ page*.
* **Section 2 – [Personas description]**. Personas should be real people, but at the same time they should be representative of the role they represent. The personas should be extracted from the scenario above, with the following information: *name*, *age*, *interest*, *application usage* and *short description*. There should *be max 20 lines of description per persona*. This part should make explicit what is already implicit in the scenario description namely, their role in life, why they should be interested in the system, and in which modalities they would like to use it.
* **Section 3 – [Storytelling definition]**.This part should be the breakdown/expansion of the scenario described above. It should be an itemized sequence of steps with three components: *i)* the persona; *ii)* the real world action/goal of the persona; *iii)* the corresponding action by the system/demo which support the persona in doing what she needs to do. *This section should not be longer than 3 pages (and not be shorter than 2 pages).*
* **Section 4 – [Generalized queries]**. In this section the actions listed in the section above should be formalized into (generalized) queries. For queries we suggest to create a second table which reports how queries can be generalized from the actions in the plot above. *This section should not be longer than 2 pages (and not be shorter than 1 page).*
* **Section 5 – [Model design]**. This section provides the diagram that represents the EER model (by using yEd) or the relational model (by using Workbench) along with the related file(s). The model should be extracted from the queries above and organized in three levels of decreasing importance: *i)* core entity types (plus relations and attributes); *ii)* auxiliary entity types (plus relations and attributes); *iii)* common entity types (plus relations and attributes) (e.g, space, time). A motivation for the selected core, auxiliary and common etypes should be provided, along with an explanation of what datasets information have been used. *This section should not be shorter than 2 pages and no longer than 3 pages.*
* **Section 6 – [Related ontology proposal and (optionally) first formalization attempt]**.Considering the model, provide a list of (maximum 5 and minimum 3) ontologies that can be used for as starting point for formalizing that kind of information. Make a ranking of the ontologies and motivate the choices. As optional task make a first formal model by using Protégé. *This section should not be longer than 3 pages (and not be shorter than 2 pages).*
* **Section 7** – **[Final considerations and open issues]**.

# Domain (Formal) Modelling (difficulty: “Medium/High”)

# *Assignment Structure*

* **Input** 
  + Topic and/or Tech Report(s) and/or Informal Model
  + CSK
* **Tool(s)**
  + yED
  + WordNet
  + Protégé
* **Output**
  + Tech report (the structure is described below, sub-section “c”)
  + Protégé File

# *Process Description*

1. **Select a topic and/or tech report and/or informal model.** Notice that the input *i)* can be provided by the course designers or *ii)* can be provided by the students. In the second case a proposal should be submitted to the course designers for approval.
2. **Formalize the model.** For this task the use of Protégé is mandatory. A specific template for addressing this task will be provided and explained during the classes.
3. **Upload lexical information.** For this task the use of WordNet is mandatory. The information will be uploaded according to the methodology we explain during the classes.
4. **Map the domain formal model with the top-level.** For this task the use of the CSK is mandatory.
5. **Generate the technical report.** (Guidelines below, sub-section “c”).

# *Tech Report Description*

* **Section 1 – [Scenario description]**. The scenario is a description of the situation that tells how the storytelling will be developed. It is in plain text. It is like telling a story to a kid describing a real world scenario with real needs, real people, real action. People should have names, an age, gender specified, roles, and should take action. *The scenario description should not exceed the maximum of a page and be at least ½ page*.
* **Section 2 – [Model formalization description]**. This section provides the design of the domain ontology along with the provision of the related file(s). The ontology must be created according to the methodology explained during the classes. A full list of classes, properties and datatypes should be provided, along with a description of the process followed for the creation of the ontology. *This section should not be shorter than 2 pages and no longer than 3 pages.*
* **Section 3 – [Lexical information upload description]**. In this section the lexical information uploading task will be described along with a full list of the synsets used for the task and a description of the motivations grounding the choices. *This section should not be longer than 2 pages (and not be shorter than 1 page and half).*
* **Section 4 – [Top-level Grounding]**.This section describes the mapping to the CSK, i.e., the top-level ontology provided and explained during the classes. The top-level grounding section should provide a diagram that represents the output of the mapping (domain roots and top-level leaves), along with a description of some issues rised by the mapping process (for the diagram the use of yEd is mandatory. *This section should not exceed the maximum of 2 page and be at least 1 page*.
* **Section 5 – [Model visualization]**.This section describes a full visualization of the output model. Here, for the generation of the diagram, yEd or Protégé can be used. *This section should not exceed the maximum of 2 page and be at least 1 page*.
* **Section 6 – [Generalized queries]**.Starting from the formal model, formalize a complete list of queries with example results.*This section should not be longer than 2 pages (and not be shorter than 1 page).*
* **Section 7** **– [Final considerations and open issues]**.

# Data Integration (difficulty: “High”)

# *Assignment Structure*

* **Input** 
  + Dataset(s)
  + Owl file
* **Tool(s)**
  + Proégé
  + Karma
  + RapidMiner
* **Output**
  + Karma file
  + Tech report (the structure is described below, sub-section “c”)

# *Process Description*

1. **Select dataset(s) and ontology.** Notice that the inputs *i)* can be provided by the course designers only.
2. **Upload the owl file.** This task is addressed by using KARMA.
3. **Upload and map dataset(s).** This task is addressed by using KARMA.
4. **Analyze the output data.** For this task the use of RapidMiner is mandatory. Notice that we allow the students to choose how to analyze the data by selecting some of the techniques explained during the classes.
5. **Generate the technical report.** (Guidelines below, sub-section “c”).

# *Tech Report Description*

* **Section 1 – [Input Dataset(s) description]**. This section should provide a short description of the datasets and a short motivation grounding the need of their integration. *This section should not exceed the maximum of a page and be at least ½ page.*
* **Section 2 – [ER Model design and Ontology description]**. This section should provide an ER model (using yEd) and corresponding description that can be used for representing the input ontology and related data. Moreover, this section should provide a detailed description of the input ontology, with consistency check (by using Protégé), the list of classes, properties and datatypes, along with a visualization of it and the motivation for its usage in relation to the reference input datasets.*The ontology description should not exceed the maximum of 4 pages and be at least of 3 pages*.
* **Section 3 – [Integration process description]**. This section should provide a detailed description of the process followed by using KARMA, by explaining the generated issues and the gap between expectations and output. *This section should not exceed the maximum of 3 pages and be at least of 2 pages*.
* **Section 4 – [Output Dataset(s) and queries description]**. This section should provide a short description of the output datasets along with a complete list of queries with example results (taken from the dataset).This section should not be longer than 2 pages (and not be shorter than 1 page). *This section should not exceed the maximum of 3 pages and be at least of 2 pages*.
* **Section 5** **– [Input/Output Dataset(s) comparative analysis description]**. Here a comparative description of the dataset along with a detailed (comparative) analysis should be provided (number of queries, attributes, instances, etc.) *This section should not exceed the maximum of 4 pages and be at least of 3 pages*.
* **Section 6** **– [DB generation proposal]**. Here a description of the process by which a DB can be generated from the output data should be provided. *This section should not exceed the maximum of 2 pages and be at least 1 page.*
* **Section 7** **– [Final considerations and open issues]**.

1. This is how a footnote should appear. [↑](#footnote-ref-1)
2. Note the line separating the footnotes from the text. [↑](#footnote-ref-2)