Barriers and challenges of using medical coding systems
Surján, G.

Citation for published version (APA):
Chapter VII. A PILOT ONTOLOGICAL MODEL OF PUBLIC HEALTH INDICATORS

Abstract
There are various public health databases in the world aiming to provide data to compare health conditions in different countries. Their data sets are more or less overlapping but data from different databases and different countries are hard to compare due to different definitions and interpretations. Our aim was to create a core ontological model that is able to represent public health indicators. We assumed that by such representation comparability and quality of data could be improved.

METHOD Three sets of indicators were taken, and a core ontology was built from information objects describing their top level entities. The Protégé ontology editor with RDF backend was used for building the ontology. The used indicator sets were the indicators of the Health for All Database of the World Health Organisation (HFA), the OECD Health Data, and the set of indicators proposed by the European Community Health Indicators (ECHI) European project. Then 19 indicators selected from HFA were represented using the core ontology. Strength and weaknesses of the descriptive capability of the model were studied.

RESULT The drafted core model seems to be useful in representing many of the public health indicators. In some cases it really helps improve comparability. However some of the semantic details can not be sufficiently expressed by the used ontology representation language. There is a need of merging other domain ontologies to represent indicators related to other domains, such as economy, social and environmental sciences.

1. Introduction
There is no effective health policy without information on the health status of the society. To support decision makers at a national or even international level various databases became available (either on the net or offline) containing information on public health, health care systems and health related indicators, like demographic, socio-economic, and environmental data. The best known examples in Europe are the OECD health data (OECD HD, www.oecd.org/health/healthdata) and the health for all database of the WHO (HFA http://www.who.dk/hfadb). In the United States CDC (Centers for Disease Control and Prevention) provides a lot of such information. (http://www.cdc.gov). The conceptual structure of the OECD and WHO databases are more or less similar but their data are not comparable in most of the cases due to many differences in the details. There are several projects all over the world [1], [4] aiming to define a standard set of indicators. The standardisation is obviously important for comparative studies in different populations. Since the differences between the indicators are coming from different needs we do not believe that all actors in the scene will ever agree on one single model. As a central institute for health information in Hungary we also have to serve the needs of
various bodies and groups, and we also have to ensure data consistency. In this study our goal was to see whether it is possible and helpful to represent health indicators in an ontology, and use it as a tool to achieve coherence among different databases serving various actors and purposes.

2. Overview of the field

2.1. Public health indicators

In public health community indicators are considered as data that in some way characterise health status and health influencing conditions of a population. Many of them are numeric data resulting from either enumeration of observed entities or from measuring some quantitative property (number of deaths, number of disease incidents, expenditure on drugs, etc.). Some experts argue that these data are not informative in themselves, and suggest to use the term indicator only if these numbers are normalised in some way, e.g. "death rate", i.e. the number of deaths divided by the number of people in a population, etc. We do not want to enter this discussion. From our point of view these are two different kinds of indicators. In terms of this paper, we use the term 'absolute indicator' for data resulting directly from observations and 'relative indicator' for data which are calculated from other (either direct or indirect) indicators. Relative indicators are defined entirely by their calculation formula. A precise ontological definition is necessary for absolute indicators. For example, we have to know precisely what "number of deaths" and "number of people in a population" mean. If these are defined, then the "death rate" (the number of deaths per 100 000 of population) does not require further ontological definition.

Although absolute indicators are obviously less useful for cross-country comparisons, existing public health databases use both absolute and relative indicators. The structural organisation, in which these databases present their data are different, e.g. the HFA database provides a hierarchichal indicator tree, and each indicator has two dimensions: location (countries) and time (years). For instance "Standardised death rate due to lung cancer in all ages" and "Standardises death rate due to lung cancer in age between 0-64 years" are both indicators on the second level of the indicator tree. The latter is subdivided on the third level into male and female, while the former is subdivided into five categories: male, female, age over 65 years, male, female and both sexes. The structure is shown in Figure VII-1. The user can select any of the items marked with the ☐ symbol, then the countries and years of interest have to be specified before retrieving data. In this structure "lung cancer in all ages" subsumes "lung cancer in age of 65 and above" but does not subsume "lung cancer between 0 and 64 years".
MORTALITY BASED INDICATORS

- Standardised death rate trachea/bronchus/lung cancer 0-64 years per 100 000 citizens
- Standardised death rate trachea/bronchus/lung cancer 0-64 years per 100 000 citizens male
- Standardised death rate trachea/bronchus/lung cancer 0-64 years per 100 000 citizens female
- Standardised death rate trachea/bronchus/lung cancer all ages per 100 000 citizens
- Standardised death rate trachea/bronchus/lung cancer all ages per 100 000 citizens male
- Standardised death rate trachea/bronchus/lung cancer all ages per 100 000 citizens female
- Standardised death rate trachea/bronchus/lung cancer 65+ years per 100 000 citizens
- Standardised death rate trachea/bronchus/lung cancer 65+ years per 100 000 citizens male
- Standardised death rate trachea/bronchus/lung cancer 65+ years per 100 000 citizens female

Figure VII-1 Part of the HFA structure of indicators

In the OECD HD the selection of an indicator requires three steps. First the user has to choose a "Chapter" from a three level tree hierarchy. Only the leaves of the tree (i.e. the third level items) can be selected. The following step is the selection of items from a plain list called "Variable". The list of the available variables depends on the selected "chapter" so the variables can be considered as the fourth level of the tree. The last step is the selection of the "Unit". It is again a list, depending on the selected chapter, but – with a few reasonable exceptions – does not depend on the selected variable. (For example gender is a unit for cancer incidence. An example of an exception is that you can not choose 'male' for cancer of female genital organs.)

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>VARIABLE</th>
<th>UNIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health status</td>
<td>Malignant neoplasms</td>
<td>Number of female deaths</td>
</tr>
<tr>
<td></td>
<td>Malignant neoplasms of the colon</td>
<td>Number of male deaths</td>
</tr>
<tr>
<td></td>
<td>Malignant neoplasms of the lung</td>
<td>Number of total deaths</td>
</tr>
<tr>
<td></td>
<td>Malignant neoplasms of the female breast</td>
<td>Deaths per 100 000 female (standardised)</td>
</tr>
<tr>
<td></td>
<td>Malignant neoplasms of the cervix</td>
<td>Deaths per 100 000 male (standardised)</td>
</tr>
<tr>
<td></td>
<td>Malignant neoplasms of the prostate</td>
<td>Deaths per 100 000 male (standardised)</td>
</tr>
<tr>
<td>Health status</td>
<td>Number of female cases</td>
<td>Number of female cases</td>
</tr>
<tr>
<td></td>
<td>Number of male cases</td>
<td>Number of male cases</td>
</tr>
<tr>
<td></td>
<td>Number of total cases</td>
<td>Number of total cases</td>
</tr>
<tr>
<td></td>
<td>Number of female cases</td>
<td>Number of female cases</td>
</tr>
<tr>
<td></td>
<td>Number of male cases</td>
<td>Number of male cases</td>
</tr>
<tr>
<td></td>
<td>Number of total cases</td>
<td>Number of total cases</td>
</tr>
<tr>
<td>Mortality</td>
<td>Number of female cases</td>
<td>Number of female cases</td>
</tr>
<tr>
<td>Causes of mortality</td>
<td>Number of male cases</td>
<td>Number of male cases</td>
</tr>
<tr>
<td>Life expectancy</td>
<td>Number of total cases</td>
<td>Number of total cases</td>
</tr>
<tr>
<td></td>
<td>Number of female cases</td>
<td>Number of female cases</td>
</tr>
<tr>
<td></td>
<td>Number of male cases</td>
<td>Number of male cases</td>
</tr>
<tr>
<td></td>
<td>Number of total cases</td>
<td>Number of total cases</td>
</tr>
<tr>
<td>Perceived health status</td>
<td>Number of female cases</td>
<td>Number of female cases</td>
</tr>
<tr>
<td>Infant health</td>
<td>Number of male cases</td>
<td>Number of male cases</td>
</tr>
<tr>
<td></td>
<td>Number of total cases</td>
<td>Number of total cases</td>
</tr>
<tr>
<td>Causes</td>
<td>Number of female cases</td>
<td>Number of female cases</td>
</tr>
<tr>
<td>Cancer</td>
<td>Number of male cases</td>
<td>Number of male cases</td>
</tr>
<tr>
<td></td>
<td>Number of total cases</td>
<td>Number of total cases</td>
</tr>
<tr>
<td>Injuries</td>
<td>Number of female cases</td>
<td>Number of female cases</td>
</tr>
<tr>
<td></td>
<td>Number of male cases</td>
<td>Number of male cases</td>
</tr>
<tr>
<td></td>
<td>Number of total cases</td>
<td>Number of total cases</td>
</tr>
</tbody>
</table>

Figure VII-2 Part of the OECD HD structure of indicators

In Figure VII-2 we try to illustrate how variables and units depend on the choice of Chapter. Location and time are separate dimensions in the OECD HD also. Compared to HFA, this structure is more combinatorial and more consistent. However the structure is still not quite coherent. E.g. "Cancer" is a chapter in morbidity, while "malignant neoplasm" (a synonym for cancer) is a variable in mortality.
Both HFA and OECD HD structures seem to be arbitrary. Retrieving identical or comparable data from the two databases is difficult. None of them seems to be based on a well-defined conceptual model. For example, in the HFA database "Salmonellosis cases" (the number of Salmonella infections) is located in the "Environment" group, and not – as one could expect – in the morbidity group. Similarly "Standardised death rates due to appendicitis" is in "Health care utilisation and expenditure", and not in the mortality based indicators. The result of this arbitrary structure is that the reuse of these data is difficult and possible only manually.

Based on these experiences we started to develop a conceptual framework, the so-called IDA model, which allows us to arrange public health data into a space consisting of a hierarchical tree of the observed phenomena and a set of dimensions, corresponding to the observable properties of the phenomena. The model is described in more detail in a previous publication [5]. Here we restrict the description to an overview of its logical structure.

### 3. The IDA model

#### 3.1. Observable entities and their properties
Different actors in the health care system observe various phenomena, e.g. physicians observe patients having certain diseases, administrators "observe" how many beds they have in their hospital, etc. These actors are observers who provide some sort of reports and send them to primary data collecting organisations (insurance agencies, various registries etc.). Primary data collecting organisations provide statistics on their data by aggregations according to various properties of the things. For example, it is possible to determine how many women die from lung cancer in a certain part of the country, or the number of newborns below a certain body weight etc. In the IDA model these properties of the observed things (gender, cause of death, body weight etc.) form dimensions.

#### 3.2. Conceptual hierarchy of the entities
Generally users of public health databases have to navigate through several hundred indicators. To assist this navigation, indicators are usually grouped into a limited number of categories, like "health status", "health-care resources", "health influencing factors". This categorisation is arbitrary, and the categories are not necessarily disjoint. It can be hard to decide whether something is a resource or a factor influencing health status, since a health care resource is used because it has some influence on health. These categories are about the purpose of things rather than about their nature. On one side this is reasonable, because it is useless to collect data on health without purpose. Public health indicators are used to monitor (measure) and improve the health status of the population. Since health is a rather complex thing, indicators are used as building elements of a mosaic. In that sense, each indicator must have a purpose: this is the aspect of health which they describe. Cigarette consumption for example measures the risk of a heart attack or lung cancer (among others). From this point of view the purpose of this information is to estimate the risk. On the other hand, one may say, that smoking
is a behavioural pattern, which indicates psychological disturbances. Neurotic people are likely to smoke much more than psychologically well balanced people. In that sense the amount of cigarette consumption is an indicator of health status. This is another purpose of the same information. It is important to emphasise, that the relation between cigarette smoking and either risk factors or health status descriptors is not an <IsA> relation. If for some unexpected reason smoking never would cause any harm in the future, it still would remain smoking but would not remain a risk factor. One indicator might have different purposes, while different indicators might have the same purpose. For example, number of traffic accidents with lethal outcome, and number of deaths due to traffic accidents are obviously different ontological entities and also different in their number but both of them have the same purpose in public health: to measure how seriously traffic threatens our lives. The IDA model is based on the ontological hierarchy of the things and not on their purposes in public health.

3.3. Dimensions

In the IDA model the attributes (the properties) of each observed entity are represented in dimensions. In other words, to each entity belong a set of dimensions, corresponding to the observed properties of the entity. As shown, the well-known public health databases use two dimensions: location and time. These dimensions are universal: it is hard to imagine to observe something which does not occur in a given part of space and time. Other properties, like gender, age, disease are not universal: hospital beds have neither gender, nor disease, birth has no age etc. In the studied databases such things are represented in the indicator tree. As we saw for instance, in the HFA database "Standardised death rate trachea/bronchus/lung cancer all ages per 100 000 citizens male" and "Standardised death rate trachea/bronchus/lung cancer all ages per 100 000 citizens female" are two different indicators, both of them are subsumed by "Standardised death rate trachea/bronchus/lung cancer all ages per 100 000 citizens" In the IDA model we have only one entity for these three indicators, this is "death". This entity has the following dimensions: "Disease" as cause of death, "age", "gender", and a so called "measure" which describes how we express the amount of the observed thing. In this case it is "rate per 100 000 citizens, standardised to the European population". So the IDA representation of the above mentioned three indicators is illustrated in Table VII-1.
Chapter VII – Ontology of public health indicators

Table VII-1 The IDA representation of indicators

<table>
<thead>
<tr>
<th>Entity</th>
<th>Dimensions</th>
<th>Gender</th>
<th>Age</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>Trachea/bronchus/lung cancer</td>
<td>Both</td>
<td>All</td>
<td>Rate per 100 000 standardised</td>
</tr>
<tr>
<td>Death</td>
<td>Trachea/bronchus/lung cancer</td>
<td>Male</td>
<td>All</td>
<td>Rate per 100 000 standardised</td>
</tr>
<tr>
<td>Death</td>
<td>Trachea/bronchus/lung cancer</td>
<td>Female</td>
<td>All</td>
<td>Rate per 100 000 standardised</td>
</tr>
</tbody>
</table>

(For simplicity, universal dimensions, space and time, are missing from the illustration.) By changing the value in the Disease dimension we can easily represent other mortality indicators. The benefit of this multidimensional representation is that many things share certain properties, e.g. "Disease" is a dimension pertinent to hospital admissions, morbidity related entities (incidence, prevalence) and other entities. Since we represent Disease in all these cases in the same way (as a dimension) it is easy to make comparisons between such entities. More precisely: it is possible to identify comparable entities based on the fact, that two entities can be compared to each other only if they have at least one common dimension. (It is a necessary but not sufficient condition.) The other practical reason to use dimensions instead of comprising all these things into the indicator tree is the fact, that certain dimensions have numerous possible values. For example, the Disease dimension – if we use the International Classification of Diseases – has more than 10 000 items. These items had to be repeated in the indicator tree as many times as the observed thing corresponds to a disease. Even if we would use only subsets relevant to the given topic (e.g. not all the diseases may cause death) the indicator tree could have an uncomfortably large size.

The weakness of this multidimensional model is that the semantics is not quite explicit. E.g. when we say that death is an event, which can be grouped into 'male' and 'female' death cases, we do not express what is the semantic relation between death and gender. Since death obviously has no gender, the proper statement is that a death case as an event necessarily happens with a person (we consider cases of humans only) and persons necessarily have gender. As far as users of public health databases are humans, this inaccuracy has no serious practical consequences, since humans are able to understand what is meant by "male death cases" for instance. Problems may arise however in data management. Because of the imprecise semantic representation of what data are meant, maintenance of consistency can be difficult, especially if a large amount and variety of data has to be managed. The problem may turn out to be more serious if we consider automatic data processing (e.g. data mining), where the necessary common sense human knowledge is missing. The situation is more complicated if a certain entity
Chapter VII – Ontology of public health indicators

is related to the same dimension in more than one way, e.g. 'hospital cases' are related to location both according to the location of the hospital and according to the home address of the patient. This may lead to confusion even in case of human use of data, if people are not aware of this fact.

Considering this weakness of the IDA model we decided to investigate the feasibility of a semantically more explicit ontological model of public health indicators.

4. Materials and methods

To obtain an overview of the field, top level categories of the indicators of two existing databases (HFA, and OECD HD) and the indicator set proposed by the ECHI (European Community Health Indicators) project were extracted and compared. From this set of indicator categories we tried to identify generic ontological entities and represent their attributes and relations in a core ontological model. All these entities were arranged under the top level class "PUBLIC_HEALTH_ENTITIES". In the first step we did not try to describe formally what the precise semantic relation between the entities is (e.g. whether a certain relation is an <IsA> relation or not), just to obtain an overview of which entities are related to which other ones. The next step was to define the attributes of the entities. Concerning this we restricted our attention to attributes relevant to public health issues. For example, Guarino and Welty argue that "having a brain" is an essential property for a "person" [6]. This is obvious, but practically useless for our purpose, since nobody observes the brains of the persons in the domain of public health. By defining the trivial attributes we had to extend the set of entities, e.g. "persons" trivially <HasGender> so we need an entity "gender" from which the <HasGender> attribute can take value ("male" and "female" as instances of "gender" are the allowed values of <HasGender>). The third step was to define the <IsA> hierarchies. To build up the core model the 2.1 beta version of Protégé was used. (This free ontology editor is available at http://protege.stanford.edu/.) The RDF backend was selected to represent the ontology. (Protégé is able to use various backends, like text file, RDF file, or different databases. RDF (Resource Description Framework) is an XML sublanguage commonly used to represent ontologies. See www.w3.org/RDF/ for specification). The "Ontoviz" plug-in was used to visualise selected parts of the model. A set of indicators taken from the HFA database (European Health For All database, released in 2003 June) was used to test the feasibility of the model. (This database is available for free at http://www.euro.who.int/hfadb.) Indicators were chosen from all eight top level groups of indicators (see Table VII-2). In this arbitrary choice we preferred those indicators which we felt difficult to represent in our model and skipped most of those, which apparently seemed to be easy to represent. Our goal was to identify the weak points of the core ontology. We did not calculate a measure for descriptive power (e.g. what percentage of indicators could be sufficiently represented by the model or something like that). The following 19 indicators were selected:
Demographic and socio-economic indicators
1. Mid year population
2. Annual average of inflation
3. Unemployment rate

Mortality based indicators
4. Life expectancy
5. Late neonatal cases
6. Foetal cases
7. Perinatal cases

Morbidity disability and hospital discharges
8. Incidence
9. Prevalence
10. Persons receiving social/disability benefits

Lifestyles
11. Number of regular daily smokers
12. Pure alcohol consumption

Environment
13. Sulphur dioxide emission
14. Average number of persons per room in occupied housing unit

Healthcare resources
15. % of physicians working in hospitals

Health care utilisation and Expenditure
17. Public health expenditure
18. Average length of stay acute care hospitals

Maternal and child health
19. Number of infants breast fed at 3 month of age

(In case of some relative indicators we took the corresponding absolute one, e.g. 'Acute care hospital beds' is given originally per 100 000 citizens.)

These indicators were represented in the ontology under a separate the top level class called "PUBLIC_HEALTH_INDICATORS". Then we identified relationships between these indicators and the "PUBLIC_HEALTH_ENTITIES". In other words, we represent public health indicators as specialisations of generic public health entities.

5. Results

5.1. Comparison of top level concepts

The comparison of top level categories of the health indicators is shown in Table VII-2. In this table the indicators are arranged in an order which allow us to see the similarities and dissimilarities. As the table shows four main groups could be identified, which are more or less present in all three sources. These are the following:

1. Demographic-social-economic indicators.
2. Indicators that describe health systems.
3. Health status indicators.
4. Indicators that describe health influencing environmental factors.
These groups are illustrated by different *italics* in Table VII-2.

<table>
<thead>
<tr>
<th>ECHI</th>
<th>OECD</th>
<th>HFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic and socio-economic situation</td>
<td>Economic references</td>
<td>Demographic and socio-economic indicators</td>
</tr>
<tr>
<td></td>
<td>Social protection</td>
<td></td>
</tr>
<tr>
<td>Health systems</td>
<td>Health care resources</td>
<td>Health care resources</td>
</tr>
<tr>
<td></td>
<td>Health care utilization</td>
<td>Health care utilization and Expenditure</td>
</tr>
<tr>
<td></td>
<td>Expenditures on health</td>
<td></td>
</tr>
<tr>
<td>Determinants of health</td>
<td>Financing and remuneration</td>
<td>Environment</td>
</tr>
<tr>
<td></td>
<td>Pharmaceutical market</td>
<td>Lifestyles</td>
</tr>
<tr>
<td>Health status</td>
<td>Health status</td>
<td>Morbidity Disability and Hospital discharges</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality based indicators</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maternal and child health</td>
</tr>
</tbody>
</table>

The three sources seem to be different, but we have to add that most of the indicators apparently missing from one source are included in its deeper levels. Our goal was not to match the indicator structures, just to get a good coverage about the 'things' relevant to public health.

### 5.2. The entities of the field

We tried to identify by intuition the ontological entities that are necessary to represent the above mentioned top level public health indicators, e.g. for demography we trivially need to have the entity "person". For 'health-care resources' we considered the following entities: "Organisation" "Personnel" "Medical equipment". Location and time as universal attributes were also considered. The resulting list of entities is the following:
BIRTH
DEATH
PERSON
MONEY
SOCIAL STATE
PAYMENT
SELLING/PROCUREMENT
BEHAVIOUR
HEALTH CARE RESOURCE
HEALTH CARE ORGANISATION
HEALTH CARE PERSONNEL
MEDICAL EQUIPMENT
DRUG
CHEMICAL AGENT
PHYSICAL AGENT
LIVING ORGANISM (LIVING AGENT)
DISEASE
LOCATION
TIME

5.3. The ontological model

By representing the above listed entities we got our core ontology. This is a skeleton that represents these entities as information objects. As we mentioned earlier, properties, which are essential for real entities but never occur as public health data were not considered. To make clear the difference between real objects and the information objects about them we used the symbol 'i:' before their names. (E.g. "i:BIRTH") It is also important to emphasize, that information objects are also real things, which are representations of other real things. We use the word 'skeleton' because some of the entities have no specific attributes in the core ontology, they are just placed in the taxonomy. E.g. "i:BEHAVIOUR" is an abstract entity and at this phase of the development we do not see any common attribute which necessarily holds for all entities subsumed by it. (An entity is abstract if it has no direct instances). Such entities are subjects of revision in the future.

Fitting the selected 19 public health indicators in the core ontology showed the benefits and limitations of the approach. In the next section we describe this.

5.4. Descriptive capability limitations of the model

If we consider population as an entity formed by persons living in a given geographical area, then the indicator "mid year population" is easy to represent in our ontology. (The phrase 'mid year' refers to a calculation method, which is not an ontological issue). Since here the observed entities are persons, we represent this indicator as an entity subsumed by "i:PERSON", which consequently inherits all the attributes of "i:PERSON". This may sound surprisingly, since 'person' does not subsume 'population'. But we are not speaking about persons and populations. We are speaking about "i:PERSON" as an information object used to describe persons, and "mid year population" certainly is a kind of such information objects.
The `<HasHomeAddress>` attribute is restricted to the geographical area for which the population has to be described. In other words `i:PERSON` is more generic than `i:POPULATION`, because the latter is derived from `i:PERSON` by restriction on the `<HasHomeAddress>` attribute. `i:PERSON` is used as a generic entity for definition of some other entities, e.g. for `i:HEALTH_CARE_PERSONNEL`.

Public health databases contain data which do not originate from the health care system, e.g. economic references, like "Annual average of inflation". Such indicators are hard to represent precisely in our ontology. If we consider 'annual average of inflation' as information officially announced by the Central Statistical Office, then the representation is easy but does not tell too much. If we want to represent the calculation method (which would support international comparability) then the representation would be much more difficult and out of the scope of public health. Such data do not result from an observation of events or things within the field of the health system. Representation of "Unemployment" shows a similar problem, however the observed entity (person `<HasOccupation>` unoccupied) can be easily described. The problem is, that the officially published number of unemployed refers to the so-called registered unemployment, which is different from (usually less then) the number of people who are unemployed.

In mortality based indicators the model works apparently well. "Life expectancy" is a little bit complicated at first sight, but it is nothing else than a well defined calculation based on simple mortality indicators. This indicator inherits its attributes form `i:PERSON` and `i:DEATH`. "Late neonatal death cases" can be represented by a restriction of the age of persons. "Foetal death cases" seems to be more difficult. The question, whether or not a foetus is a person, is a subject of heavy discussion. (We just refer to the serious debates on abortion.) However in our model a foetus cannot be described by the information object `i:PERSON`, because a foetus does not have the same attributes. It has gender and age but has no social state for instance. (Of course we do not want to provide arguments for the discussion of abortion with that.) For that reason we extended the core ontology with a new entity: "i:FOETUS", with the following properties: `<HasAge>`, `<HasGender>`, `<HasMother>` To represent "Perinatal death cases" is possible as a union of two kinds of events: i.e. foetal cases shortly before and newborn cases shortly after birth. Such arbitrary unions of entities are not well supported in ontologies.

Morbidity data are easy to represent, since they consist of just two main categories: incidence (number of new cases within a time period) and prevalence (number of existing cases in a population at a given time point). The ontological difference between incidence and prevalence lead us to introduce two basic abstract categories for public health entities. These are "i:EVENTS" and "i:STATES". All public health entities belong to one of these categories. Their distinction is based on the different relations of the observable entities with time. "Events" occur in a time point and they are observed over a period of time. (e.g. incidence of a disease in a given year), "States" are entities which exist over a period of time and observed at a given time point, e.g. number of people is observed at a time of a census. Our interpretation of time point and time period is coherent with the CEN pre-standard ENV 17381 (Time standards for health care specific problems). It is
Chapter VII – Ontology of public health indicators

described in [7] by Ceusters et al. According to this standard, the difference between time point and time period is pragmatic: a portion of time which has duration irrelevant or immeasurable in the given situation is considered as a time point. Both prevalence and incidence are observations of two kinds of entities: disease and person. In order to represent their difference, we had to introduce a new entity i:DISEASE_ONSET, i.e. the emerge of a new disease, while we changed the name of the entity i:DISEASE to i:DISEASE_STATE, to make clear that the former belongs to 'events' and the second to 'states'. "Persons receiving social/disability benefits", can be represented as persons, who have social state 'supported', using the i:PERSON and i:SOCIAL_STATE entities.

Among lifestyle indicators "number of regular daily smokers" is about observation of persons, who have the behaviour 'smoking' with the frequency 'daily'. At this moment we are not quite sure, if all behaviours have frequency, so this property was introduced for smoking specifically. "Pure alcohol consumption" is a different view on a similar thing. Here the observed things are not persons who drink alcohol, but a total amount of alcohol consumed. This information comes from outside of health care (from the commerce). This is similar how the OECD database measures smoking in 'grams of nicotine per capita'. They can be represented as i:SELLING/PROCUREMENT. Such data has some obvious weakness from the viewpoint of public health. Knowing how much pure alcohol is sold in a year in a country does not tell too much about alcoholism. A small number of heavy drinkers in one country and a large part of the population drinking modestly in another may consume the same amount of alcohol. (It can also happen that in a certain country people do not drink the alcohol they buy, just resell it e.g. in another country). This recognition in itself made this study worthwhile.

Also environment related indicators are again partly out of the scope of public health. Amounts or concentrations of chemical agents like "Sulphur dioxide emission", is not too difficult to represent. "Average number of persons per room in occupied housing unit" is a very interesting indicator. If we observe the number of persons and the number of 'occupied housing units' in a population, then this indicator is defined as a ratio of these two entities. For this we need to observe housing units having the attribute 'occupied'. (For instance bathrooms do not count.) But it is also possible (and very likely) that this information comes from a representative study, where interviewed persons are asked: 'How many persons do you share your room with?'. For this the entity i:PERSON should have a separate attribute. In this latter case it is also possible to study the differences of this indicator among various groups of persons (e.g. according to occupation). So the difference of the two ways is essential, and our ontological model is capable to make this very clear.

Health care resource is in our core model an abstract entity that must have a function. It has three children: 'personnel', 'health care organisation' and 'medical equipment'. "Physicians working in hospitals" can be represented easily as personnel having the function 'hospital care' and occupation 'physician'. (i:HEALTH_CARE_PERSONNEL is also subsumed by i:PERSON, so it inherits the occupation attribute from i:PERSON). Representation of "acute care hospital beds" should be done carefully. From an ontological point of view a bed has no
such a property: the same bed can be used for acute care on one day and for chronic care on the next. This means that 'acute care bed' is not <IsAKindOf>'hospital bed'. We can not speak about 'being an acute care bed' in an ontological sense. The proper representation is that 'acute care bed' is a "hospital bed", which is a "medical equipment", which has function 'acute care'. Contrary to that, 'Being a hospital bed' is not just a function of a 'bed' in general but another type of bed, since beds used in hospitals are rather different in their construction from other beds. (This does not mean of course, that a hospital bed can not be used as a bed at home.)

The two examples of Health care utilisation and expenditure are different in their nature. "Public health expenditure" is related to economy. It does not fit well in our recent core model, since it can not be considered as a selling or procurement. Financing of health services has many different methods, varying for country to country. Most of these methods strongly differ from buying something on a free market. Beyond that a definition of "public health" is also not possible within our core ontology. Anyway, "Public health expenditure" can be represented as an amount of money, which does not say too much in itself. The "average length of stay in acute care hospitals" is a calculated indicator, in which the underlying entities ('hospital treatment episode', 'hospital', 'acute care') fit the model well. A 'treatment episode' is a i:STATE in our model, since it has a relevant and measurable duration in time. This indicator inherits attributes of i:PERSON (the patient, who was treated), of i:HOSPITAL (the hospital that performed the treatment episode), of i:DISEASE (the disease the patient was treated for). Its function attribute is restricted to "acute care". This way of representation of this indicator makes it possible to study the variations of the average length of stay according to many properties (the disease, the location of the hospital, the age of the patient etc.), but still does not make clear what is the exact semantic relation between the treatment episode and the person for instance. Anyway the representation is better then it was in the IDA model, because we do not say here for instance that a hospital case has gender.

Our last examined indicator was taken from the group Maternal and child health. "Percentage of infants breast fed at 3 month of age" can be easily described by restricting the age of person to 3 month, and creating a "fed by" entity as a subtype of "behaviour". However it is not clear, whether it is a behaviour of the infant or of the mother. Beyond that this representation does not consider mother at all, which makes it impossible to study the variation of this indicator in different groups of mothers (e.g. according to age, education or occupation of the mother). Having a mother could be an attribute of i:PERSON, however in many of the cases, in which persons are observed we do not have any information about the mother. To solve the situation, we introduced two subtypes for person: i:MOTHER and i:INFANT. Of course mothers must be females and infants must have age less than one year. So this indicator inherits attributes from i:INFANT and i:MOTHER, with the restrictions on age (3 month) and "fed by" (partial breast fed or breast fed only).
6. Discussion

Ontology is in the focus of recent medical informatics research. However, we are not aware of other studies in the literature aiming at ontological representation of public health indicators. Of course there are many related works. In the United States Centers of Disease Control and Prevention (CDC) created the "Public Health Conceptual Data Model" (PHCDM), which is freely available on the web at http://www.cdc.gov/od/hsib/docs/phcdm.htm. This is an object oriented data model based on the HL7 Reference Information Model. HL7 is a widely accepted communication standard in health care. PHCDM is apparently a useful model to support data exchange among public health organisations and health care providers. Berman states, that using this model improves data quality [8]. However PHCDM does not provide an ontology of indicators and it was not designed for improvement of comparability of heterogeneous databases. Our pilot ontology of public health indicators is not intended to support data exchange. It is a step towards formal representation of public health data, to support maintenance and comparability of various public health databases, and also to support users to navigate in the expanding space of public health information.

7. Conclusions

Our aim in this study was to see, whether an ontology of public health indicators could improve comparability of various public health databases. We arrived at the following conclusions:

1. An ontology of public health indicators is a representation of information objects, i.e. data that describe observable entities in the field of health care and also of some other domains such a economy. Many of the ontologically important (essential or necessary) attributes of the observable things are not represented in the information objects which are used to describe things in health information systems. An ontology of the observed things instead of the ontology of information objects used to describe those things would be more precise but would serve our aim less.

2. The ontology of public health indicators is semantically more explicit that the previously developed multidimensional IDA model. In certain cases it was possible to represent the differences of the various data acquisition methods. (We refer here e.g. to the different methods to determine "Average number of persons per room in occupied housing unit") This at least helps to decide which data are comparable and which are not.

3. The language used to represent the ontology in this study (the RDF backend of Protégé) is not capable to express all necessary semantic features. By defining subsumption relations between information objects and indicators we can make benefit from the inheritance of the attributes. This is a major advantage of the model. Saying for instance, that mortality is an indicator subsumed by an information object used to describe persons ensures that mortality inherits person-related attributes (age, gender etc). But this still not makes explicit that we are speaking about
person who died. However the model allows us to make a difference between the person who dies and the person for instance who certified the death.

4. Ontology of abstract artefacts (by abstract artefacts we mean immaterial things created by humans), like public health data is apparently a difficult area. We can confirm Blois, who stated that artificial things are more difficult to represent than natural things [9]. The recent pilot ontological model is a compromise between the rigorous rules of formal ontology and practical applicability. We did not adopt a top level foundational ontology, (like the basic foundational ontology developed by Smith et al. [10]) because we do not know at this stage of our work, what features of a domain independent top level ontology are essential for us.

5. Public health often uses concepts of other sciences, like economy, sociology or environmental sciences. Representation of indicators based on such concepts needs incorporation of ontologies of these domains as well.

The major problem we faced during this work is that the representation language used to describe our ontology does not support the so-called role propagation. In the IDA model we had the problem, that for instance hospital treatment episodes as such have no gender, still we have to be able discriminate female and male cases. By inheritance of the attributes of information objects we are able to manage this in our recent ontological model, but – as we mentioned in the third conclusion – this still does not make clear the semantic relation between person (who has gender) and hospital treatment episode. Role propagation says that if there is relation $r_1$ which holds for $x$ and $y$, and relation $r_2$ which holds for $y$ and $z$, then $r_1$ also holds for $x$ and $z$ as the following formula says:

$$ r_1, r_2, \forall x, y, z : r_1(x, y) \land r_2(y, z) \rightarrow r_1(x, z) $$

If $r_1$ expresses the gender relation (let say $x$ means female, $y$ refers to a person) and $r_2$ expresses a "subject of treatment" relation between a person $y$ and a treatment episode $z$, then the declaration of role propagation for $r_1$ and $r_2$ could help to describe male and female cases and would make the relation between persons and treatment episodes explicit. Unfortunately the known generally used ontology description languages do not support role propagation.

We feel that the recent pilot ontology is not strong enough to be implemented in a public health database, but the ultimate goal is to build up a database which integrates data from various sources and enables us to identify comparable and incomparable data. The recent work provides some theoretical basis for this.

Acknowledgements
This work is partly supported by the Ministry of Health, Social and Family Affairs under the program "Public Health Database" and by the EU IST FW-6 507505 project "Semantic interoperability and data mining in biomedicine". The authors express their special thanks to Arie Hasman for commenting the manuscript.
Chapter VII – Ontology of public health indicators

References

[1] PG Kramers The ECHI project: health indicators for the European Community
  Eur J Public Health. 2003 Sep;13(3 Suppl):101-6


  authorities: searching for common ground. Can J Public Health. 1999 Nov-
  Dec;90 Suppl 1:S57-61

  Sante Publique. 1998 Dec;46(6):481-90

  Indicators: The IDA Model Medinfo 2004 p 1230-40

  Hyon Myaeng, (eds.), The Semantics of Relationships: An Interdisciplinary
  Perspective, Kluwer, 2002 pp. 111-126

[7] W Ceusters, F Buekens, G DeMoor, J Bernauer, I DeKeyser, G Surjan TSMI:
  a CEN/TC251 standard for time specific problems in healthcare informatics
  and telematics. Int J Med Inf 1997 Sep;46(2):87-101

[8] LE Berman, AL Fisher, L Evans, T Tilert Developing a Metadata Data Model
  for the National Health and Nutrition Examination Survey (NHANES). Proc
  AMIA Symp. 2003;789.


[10] P Grenon, B Smith, L Goldberg Biodynamic Ontology: Applying BFO in he
    Biomedical Domain in Ontologies in Medicine Ed by D Pisanelli IOS Press
    Amsterdam 2004. 20-38.