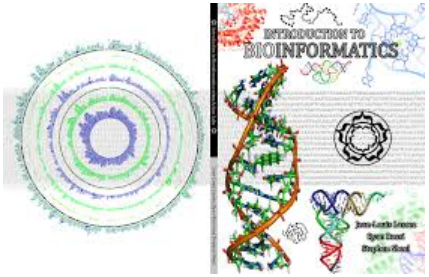
GFO-Data: Towards an Ontological Foundation of Data Semantics

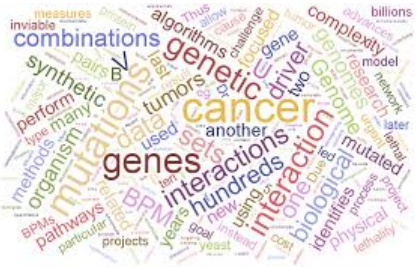
Heinrich Herre

IMISE
Research Group Onto-Med

University Leipzig

Big and Heterogeneous Data





Information Crisis 1

The *information crisis* describes the present situation caused by

- Increase of generation of information/data
- New information technologies
- Increasing of memory capacity
- Copy and transfer of data
- …

Summary: VVV Velocity, Variety, Volume

These processes lead to an information overload.

Information Crisis 2

Theses:

1. Information overload is caused by insufficient methods for abstraction and interpretation of data, and by a lack of adequate organization of knowledge.
2. Statistical correlation of data does not imply causation, and a deeper understanding of data can only achieved by/through theories or models.
3. Theories cannot be exclusively derived from data, another origin are idealizations, thought experiments, ect.
4. The onto-axiomatic method, top level ontologies, and phenotype/property ontologies are the basic means for abstraction and interpretation of data, and, furthermore, for a principled organization of knowledge

What are Data, Information and Knowledge?

The overall aim of our work, intended to be realized within a long-standing project, is to establish an ontological framework which may serve as a unifying theory of data and knowledge.

General Problem for Data

Development of a language which provides a semantic basis for data and means for their correct representation. Four subtasks can be derived.

(1)We must clarify what data are and how they can be classified (*semantic problem*),

(2)Methods to acquire these data (*acquisition problem*)

(1)Correct representation of data in a formal language (*representation problem*)

(2)Development of methods to evaluate and use these data (*utilization problem*).

What are Data, Information and Knowledge?

Engineering and Science.

Hamming (1997) "In science, if you know what you are doing, you should not be doing it. In engineering, if you do not know what you are doing, then you should not be doing it”

The current presentation is devoted to the investigation of the semantic problem for data, hence it is intended to be a contribution to the new research field of Data Science.

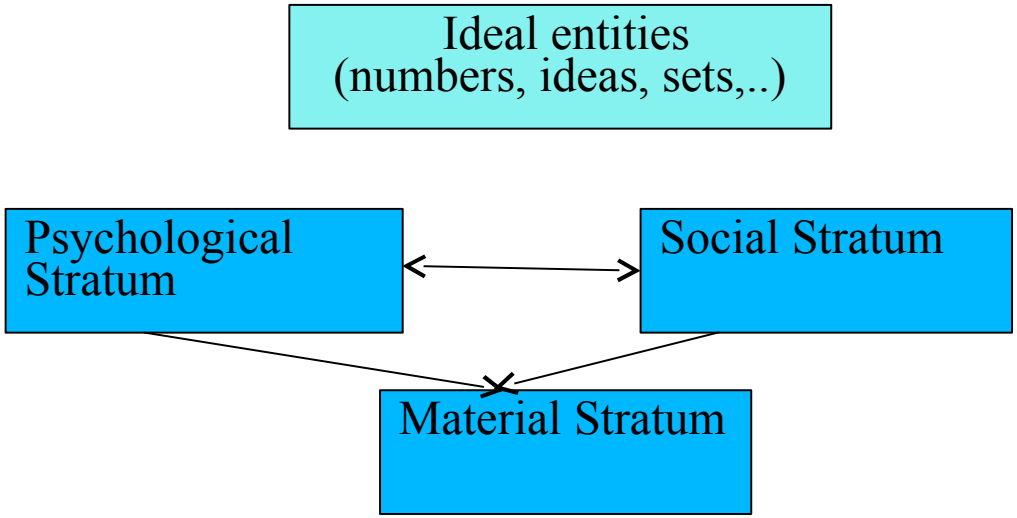
What are Data, Information and Knowledge?

The investigation is carried out within the framework of the top level ontology GFO.

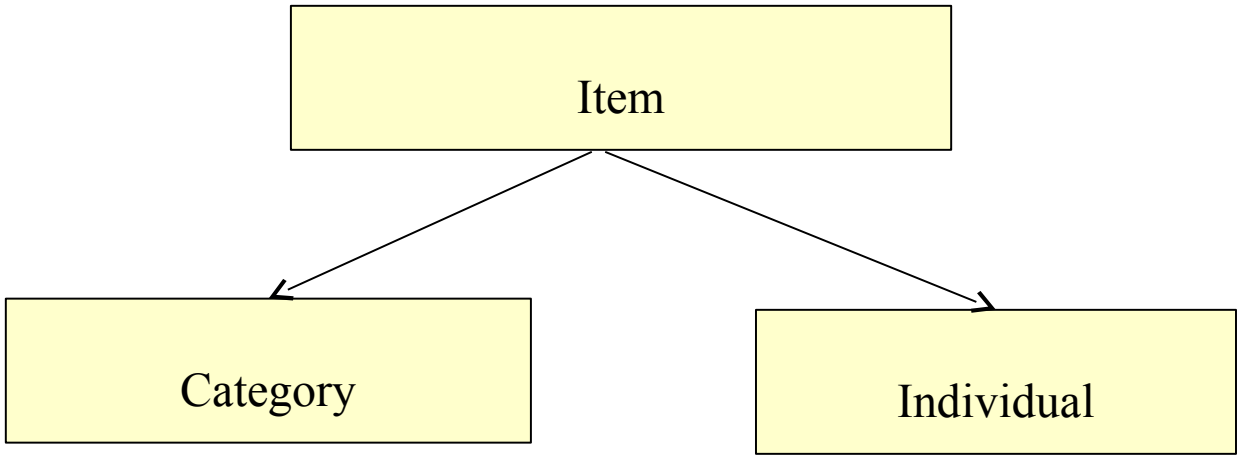
GFO (General Formal Ontology) is a foundational ontology being developed at the University of Leipzig since 1999.

GFO 1

Levels of Reality

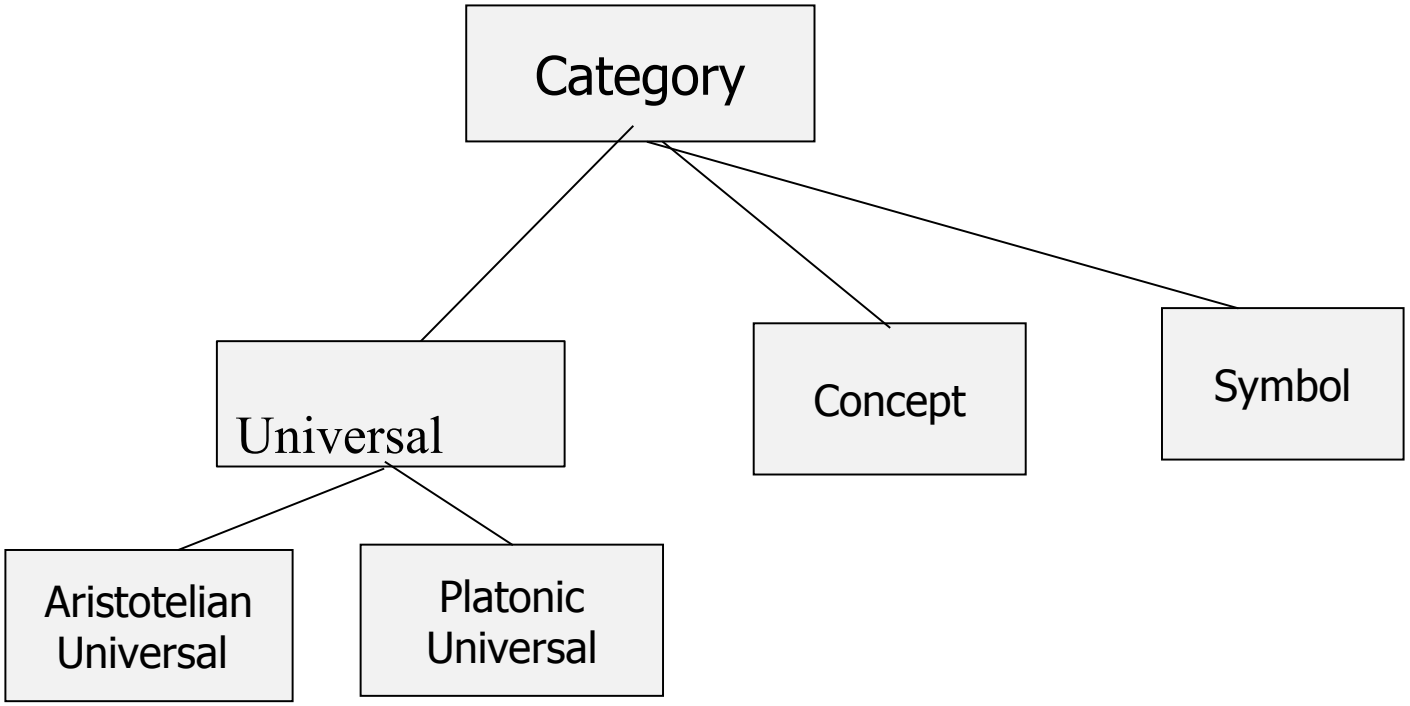


GFO 2



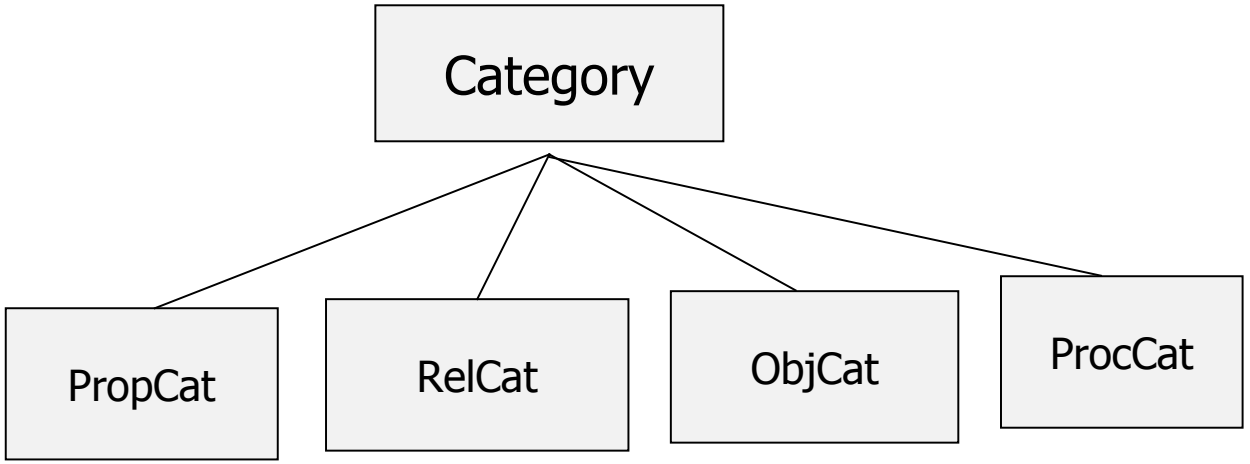
GFO 3

GFO admits various types of categories



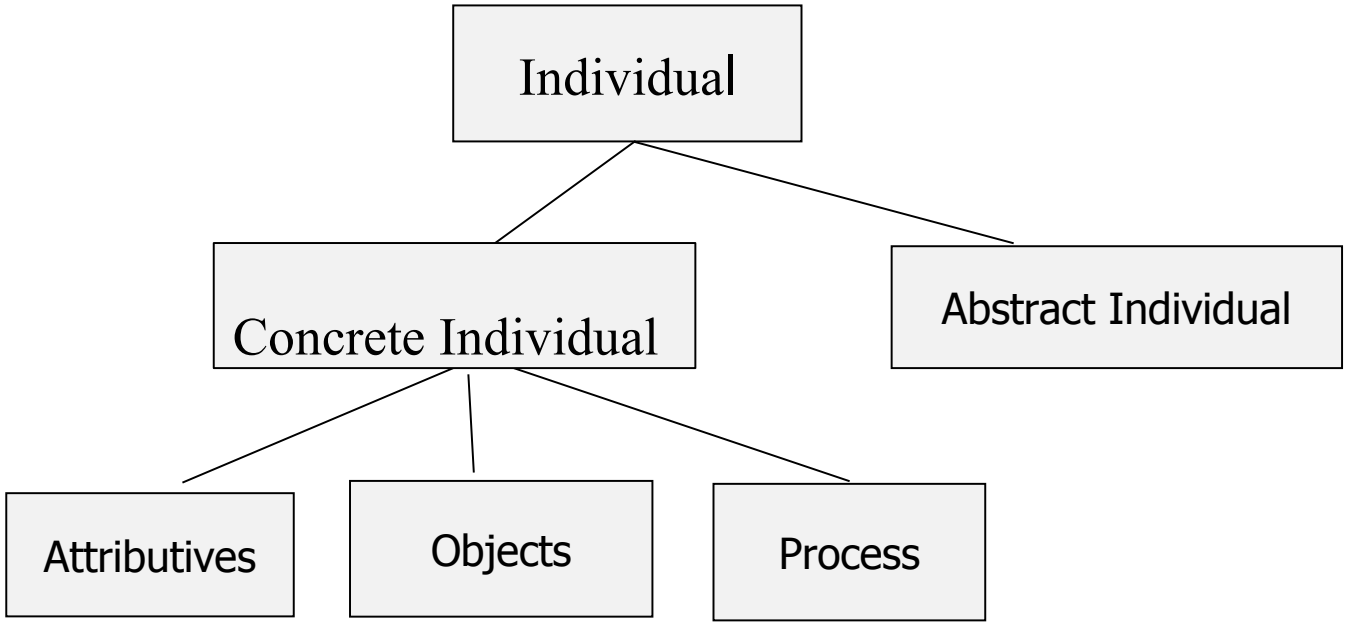
GFO 4

Types of categories: Property Categories, Relation Categories, Object Categories, Process Categories



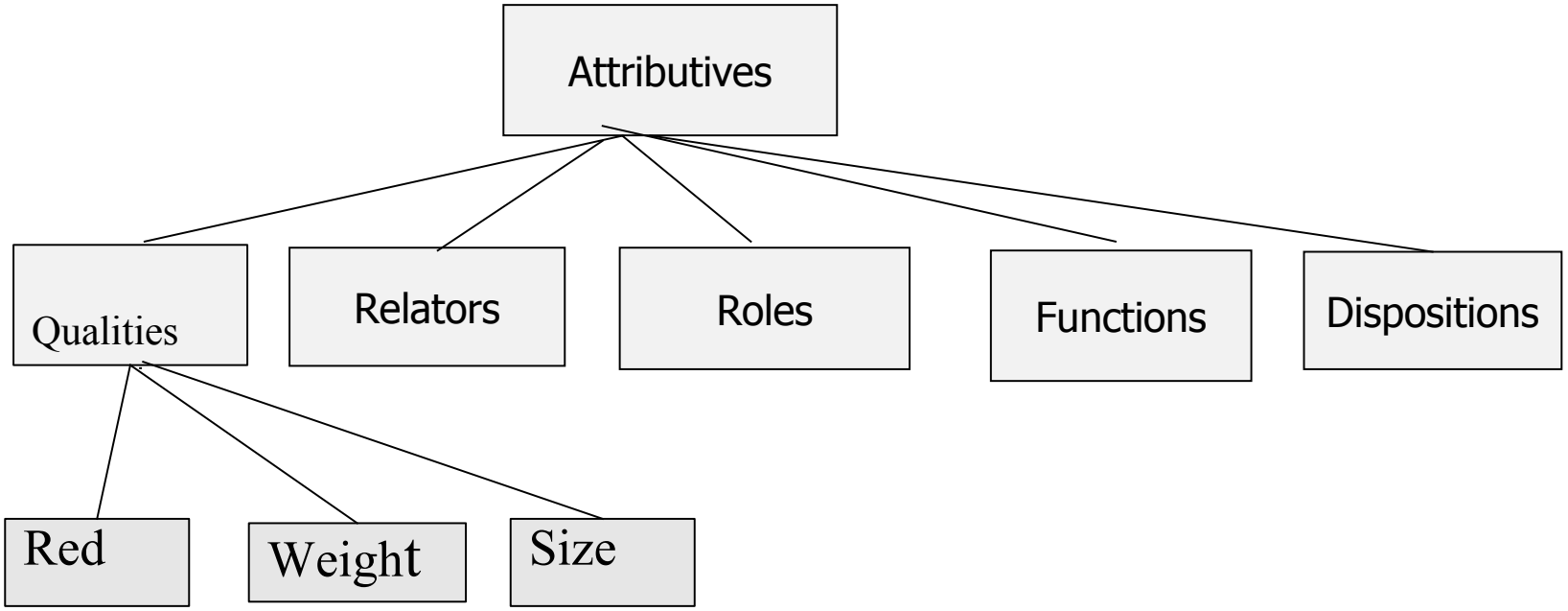
GFO 5

GFO admits various types of categories of individuals



GFO 6

GFO admits various types of attributives = data



Data 1

elementary data correspond to primitive attributives

Attributive ➔ Property, Relation, Function, Disposition, …

Attributives are existentially dependent, they need a bearer; there are various basic relations connecting attributives with their bearers.

Example: In individual red inheres in this apple (the connecting relation is the inherence relation).

Data 2

Data are classified along three dimensions:

a) Level of abstraction

b) Bearer and connecting relation

c) Complexity

Data 3

Classification w.r.t. to bearer and connecting relation

The complete specification of a property includes

a) category P,

b) binary relation R(x,y), connecting the instances of P with bearers

c) category Q of bearers

Full specification: (P, R, Q)

Example. P property Red, instances are individual Reds, Q category of material objects (say balls), R is the inherence relation

Data 4

Instances of properties are called attributives

Types of attributives: qualities, roles, relators, functions, dispositions,...

Examples

- this red of this apple, this blue of this eye (bearer is a material object being)
- the velocity of a moving body at a time point (bearer is a process)
- the pattern of a heart activity presented by a electrocardiogram (bearer is a process)

Data 5

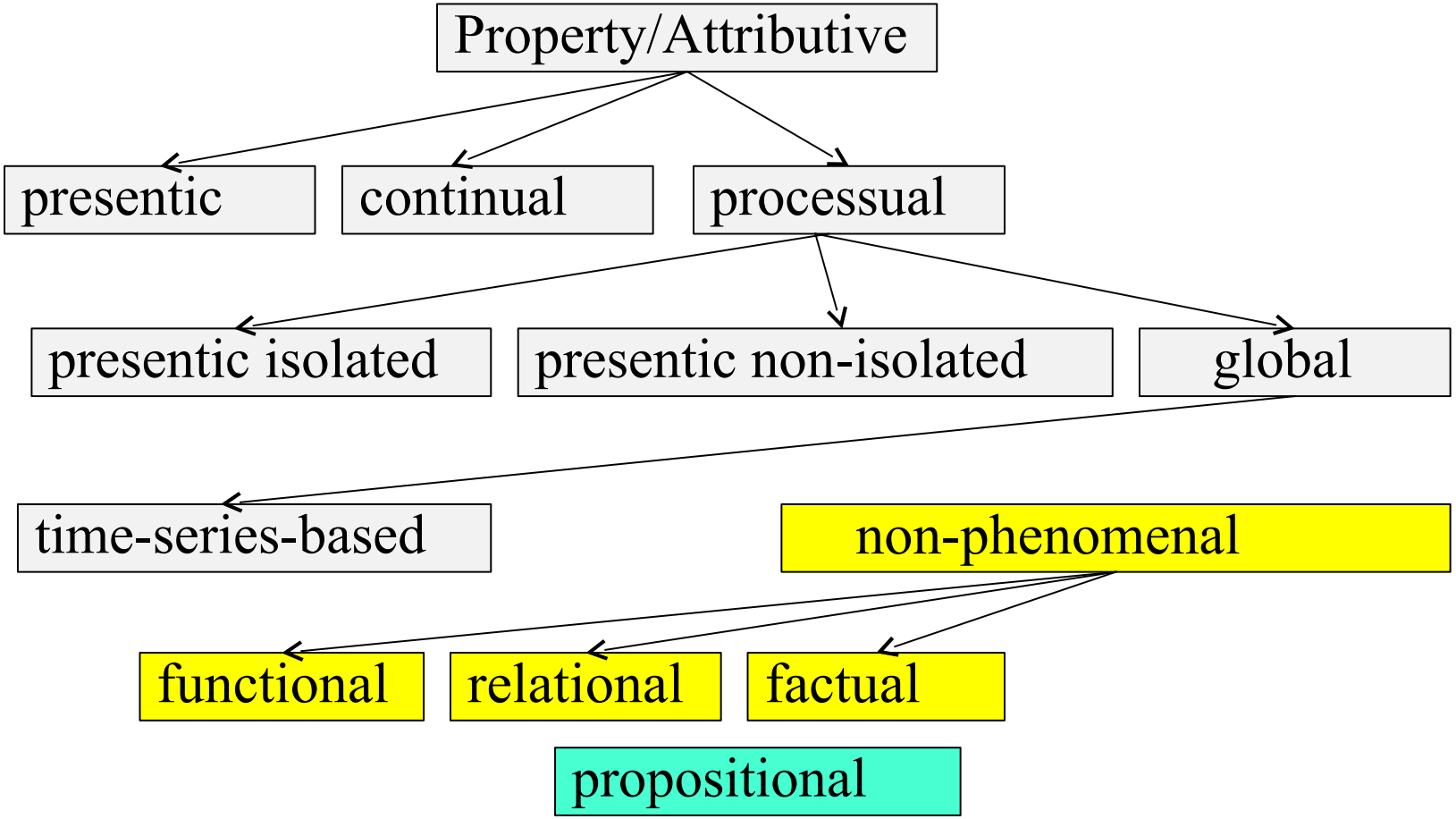
Classification w.r.t. Level of abstraction

a) Phenomenal level (sense data, data measured by instruments). The phenomenal level has various sublevels of abstraction.

b) Factual, functional, relational, situational level

c) Propositional Level

Top Level Classification of Properties



Level of abstraction 1

Phenomenal data
sense data, measurable data;
presentic, continual data, processual data (presentic, presentic non-isolated, time-series based)

Factual, relational, functional data,…
(cannot be measured and perceived by senses, for example, the function of a stone to act as a hammer)

Level of abstraction 2

The fact „John‘s drinking a beer“ is a part of reality
(not having a truth value), hence it represents factual data (factual level)

Propositional level
The proposition „John is drinking a beer“ is a proposition, and it has a truth value, hence it
exemplifies the propositional level.
The demarcation between fact and proposition
presents the distinction between data and knowledge.

Complexity of data

Elementary data
attributives having one bearer, for example,
„this red of this apple"
Description on the property level
(Red, inherence, Apple)
Instances are facts

F := <r::Red, inherence, a:Apple>

r exemplifies an elementary attributive (an elementary datum) whereas the fact F presents
a more complex datum.

Complexity of data

elementary data ➔ factual data (facts) ➔ situational data (situations)

Examples 1

Phenomenal data
a) Object-data. Data the bearer of which are material objects (also called continuants).
An object persists through time (is identical over
time), it has a life-time and is wholly present at any time-point of its life time.
„Das vollständige Dasein einer raumzeitlichen Entität zu einem Zeitpunkt ist eine Urphänomen, das uns durch die innere (reine) Anschauung gegeben ist."
Ein materielles Objekt zeigt zu jedem Zeitpunkt seiner Lebenszeit eine derartige vollständig präsente Entität, die in GFO Präsential genannt wird.

Examples 2

Properties of objects must be wholly present at time-points and objects can be understood as bundles of presentic attributives.

Examples of such presentic attributives/data:

Form (round, oval, triangular, .., i.e. geometric forms; morphology, colours, weight, etc.

Basic Task: Develop a taxonomy (ontology) of presentic properties (i.e. of object properties).

Furthermore, distinction between internal and external properties.

External (example): an object has a certain location within a coordinate system.

Examples 3

b) Processual data, i.e. attributives having a process as a bearer.
A process is a spatiotemporal entitiy which happens, which evolves through time, which has a temporal extension.
„Ein Prozess ist ein Urphänomen, das uns durch die innere Anschauung gegeben it. Dieses Urphänomen beruht auf der inneren Anschauung der Zeit."
A process can never wholly present at a time-point.
Hence, objects and processes are complementary entities.

Examples 4

Process boundaries.
The restriction of a process P to a time-point t of its temporal extension is called process boundary of P at time-point t, denoted by P(t). If P(t) is an entity it must have a presentic nature.

Integration Law of GFO. For every material object Obj there exists a process Proc(Obj) such that the process boundaries of Proc(Obj) coincide with a presentials exhibited by the object Obj.

Examples 5

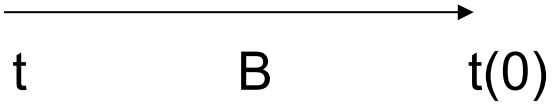
The integration law is the unique selling condition of GFO which distinguishes it from the usual 3D-Ontologies (for example BFO, DOLCE, UFO) and from the usual 4D-Ontologies (there are only processes).

Examples of processual properties.

a) Isolated presentic. A moving red ball is a process P, and at any process boundary of P there exists a presentic red ball (a presential), and the form and color of this presential is are isolated presentic properties of P. These properties are called isolated because they are determined without any process.

Examples 6

b) Presentic non-isolated.
The velocity of a moving body B at a time-point t is a non-isolated presentic processual property.
Die Bestimmung der Momentangeschwindigkeit erfordert einen Grenzübergang:
notiert v = ds/dt
Dieser Grenzübergang erfordert einen Prozess P.



Examples 7

c) processual global (bearer is a spatiotemporal process)

- Die Kreisförmigkeit der Bewegungsbahn eines sich bewegenden Körpers. Die Kreisform einer Bewegung eines Körpers ist keine präsentische Eigenschaft.
- Das Elektrokardiogramm der Herztätigkeit (als Prozess). Derartige globale Eigenschaften werden aus Zeitreihen gewonnen.
- Eine Welle eines sich ändernden Mediums (Luft, Wasser, elektromagnetisches Feld)

Examples 8

Dualität zwischen Objekt und Welle.
Wenn eine Entität E (z.B. Elementarteilchen) Eigenschaften eines Objekts Obj und einer Welle zeigt, so muss die Welle W als globale Eigenschaft des Prozesses Proc(Obj) bestrachtet werden. W kann nicht Eigenschaft von Obj sein. Objekt und Welle hängen durch das Integrationsgesetz von GFO eng zusammen.

Welche Eigenschaften besitzt der Prozess Proc(M) für ein materielles Objekt? De Broglie hat vermutet, dass jedes materielle Objekt Obj eine Materiewelle besitzt. Nach GFO muss diese Materiewelle eine Eigenschaft von Proc(Obj) sein.

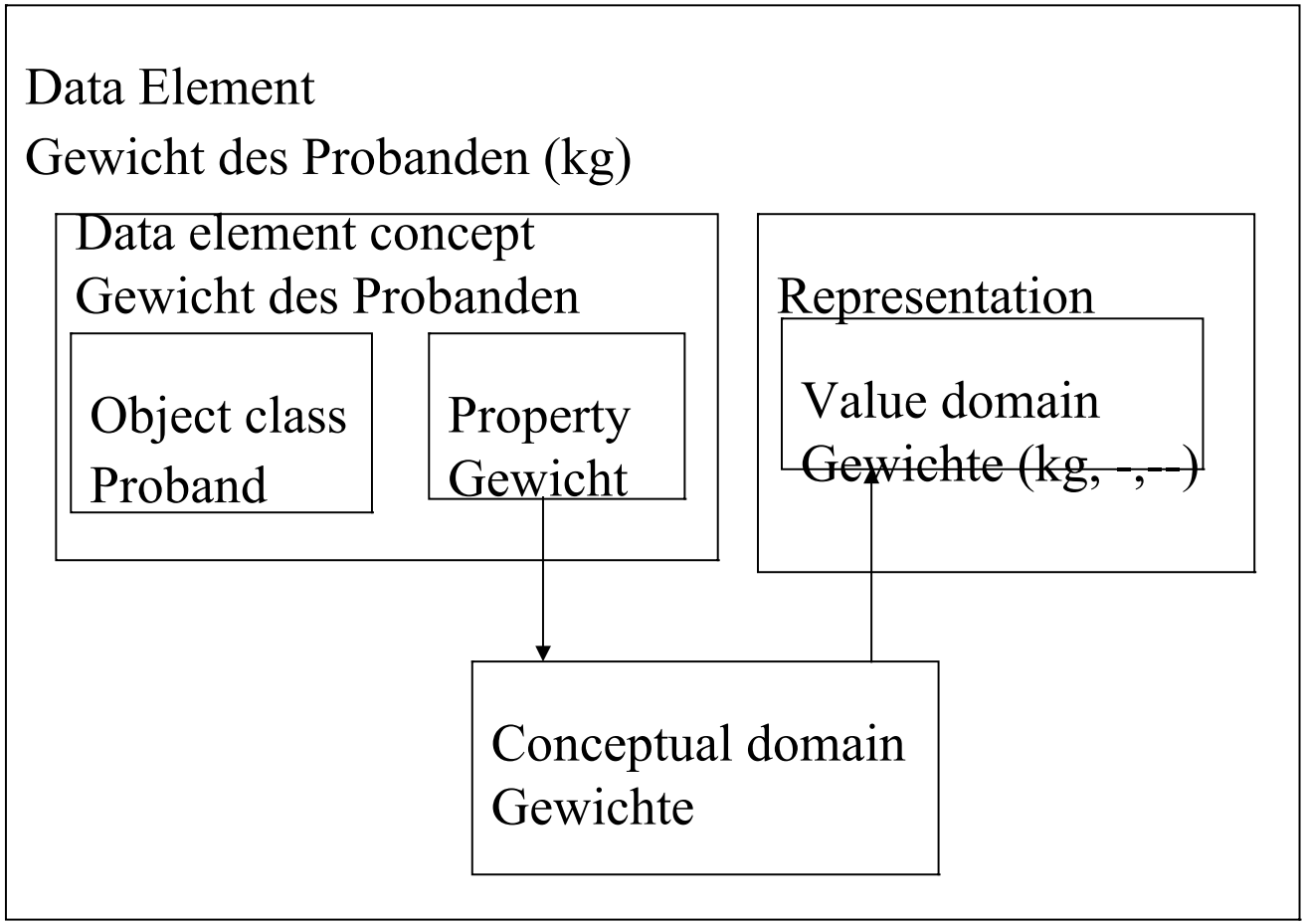
Representation of Data

Data must be represented (denoted) by symbols

Without denotation data cannot be communicated, stored, and processed. The real entities behind the symbols describe the semantics of datsa, the denotations the syntax.

GFO-Data uses an ontological foundation of the notion of data element in the ISO-standard 11179
to represent elementary data.

Representation of Properties by Use of Data Elements



Task

Various ontologies are needed:

- ontology of time and space (solved)
- ontology of data acquisition and measurement (p.s.)
- ontology of properties (p.s. ➔ PATO, HPO,..)
- ontology of relations, roles, and functions (p.s.)
- ontology of facts, propositions, and situations (p.s.)
- ontology of information entities (p.s.)

Applications

The GFO-Data Framework can be used for semantic data integration in various fields, in particular in

- Bioinformatics
- Biomedicine
- Clinical Trials

In all these fields there are already practical applications.

Data, Concepts and Knowledge

Information, according to the GFO-method (Onto-axiomatic method), is organized as follows:

I. Data ➔ Concepts ➔ Knowledge (propositions)
(levels of abstraction between data and knowledge)

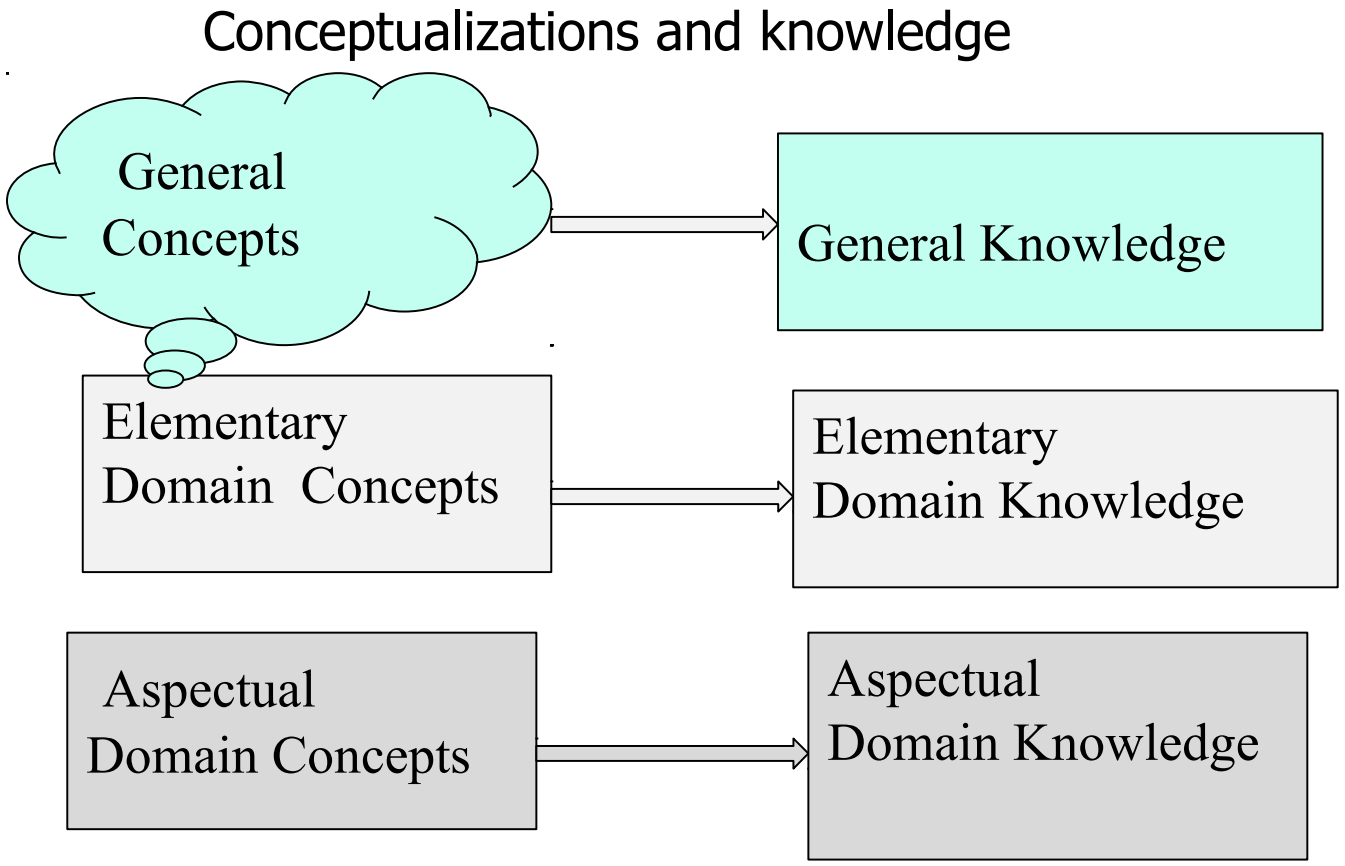
II. Organization of Knowledge
(levels of abstraction within the knowledge level)

Domain-knowledge ➔ Upper Domain Knowledge ➔

Top Level Knowledge

Domain Knowledge = elementary and aspectual knowledge

Levels of Knowledge



End

Thank you!