# Information Process Produce the Imperfections in Data

Advanced Draft (#4345 Version 8af - 7.647 Words) with incomplete references

*Andrew U. Frank*

*Department of Geoinformation and Cartography,*

*Technical University Vienna,*

*Gusshausstrasse 27-29/127-1*

*A-1040 Vienna, Austria*

*frank@geoinfo.tuwien.ac.at*

**Abstract:** All knowledge derives from observation, which is refined and restructured in complex information processes. This article analyzes the information processes with which data or knowledge are transformed from observation to compact, abstract knowledge and decisions. Three ontological tiers for data or knowledge are differentiated:

- observations of physical properties at a point (sense data);
- formation of object data with summary descriptive properties (granulation) and their mental classification;
- conceptual constructions with representations in context that can be communicated.

It derives the properties of the imperfections in data from the properties of the information processes and argues that all imperfections in our knowledge must be explained from the properties of the information processes involved. Observations are mainly influenced by random errors that can be modeled by a normal distribution. Granulation can be described by transformations of probability distribution functions (PDF) and mental classification results in fuzzy values. Constructions are free of error within the defining context; better models than supervaluation for the imperfections introduced by change of context are critically needed for semantic data interoperability.

# 1 Introduction

The goal of this article is to show the connection between the processes with which we produce and use data and the imperfections in these data (the argument is developed at a high abstraction level where data and knowledge are not distinguished; I will mostly use the neutral term data). The hypothesis that all imperfections in the data or knowledge are the result of imperfections in the information processes has two practical implications:

1. Imperfections in information processes must be analyzed to identify the types of imperfections they introduce into the data.
2. Theories about imperfections in data and methods to reduce imperfections must be justified by empirically observable properties of information processes.

I will speak of imperfections in the data, including all consequences of the use of the data, which may eventually lead to non-optimal decisions. This viewpoint links data to decisions and assesses its suitability for a decision—the "fitness for use" paradigm (Chrisman 1989)—and avoids the broad and unspecific concept of data quality.

The imperfections in the data are the result of properties of the processes that are used to observe reality, to transform the data from one form to another, and eventually to use it to make a decision.

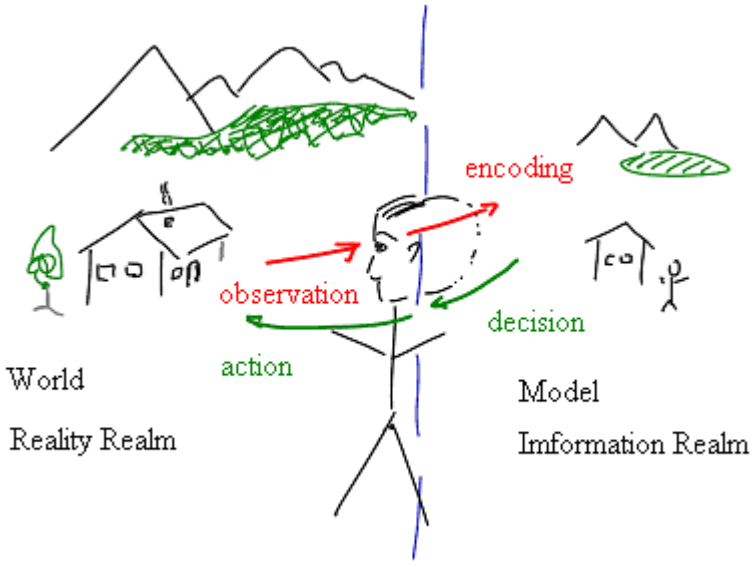


Figure 1: The connection between physical reality and data

Description of the imperfections must lead to a theory that links observable imperfections to empirically justified aspects of the information processes. The approach I use extends a realistic ontology and epistemology that broadly represents what I think is a consensus of what is implemented in many GIS and used implicitly by many GIScience researchers (Frank 2001; Frank 2003). The ontology is first extended to include the information processes and then the properties of these processes are linked to the imperfections observed in the data (Figure 2). A connection between GIS ontology and spatial data quality research is established.

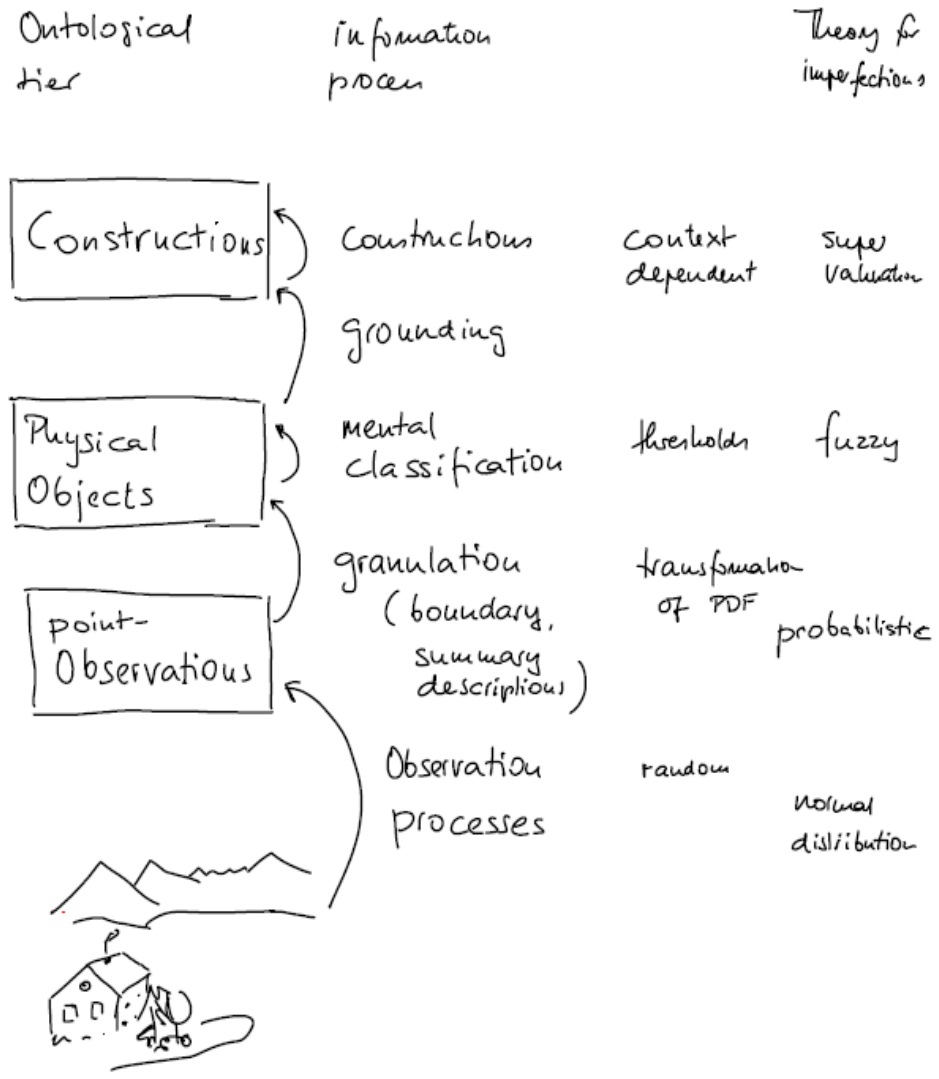


Figure 2: Tiers of ontology and information processes transforming data between them

This article excludes the very important issue of scale effects, which are present at all levels of information processing—from instant field of vision [] in observation to the modifiable area unit

problem (MAUP (Openshaw, Charlton et al. 1987)). This article also does not discuss processes in reality, which could better justify some arguments [Kuhn COSIT 2007]. Both restrictions are necessary to keep focus, and are also caused by a paucity of previous results useful in this context.

## 2 Terminology

*Imperfection* is the general term that is used here to describe the deviation from the broadly generalized, naïve concept of "truth". Perfect data are conceptually a perfect representation of reality and leads to perfect (optimal) decisions. Comparing data to "ground truth" allows us to determine the quality of the data—at least this seems to be the naïve assumption behind concepts like "ground truth" (Pickles 1995) and "data quality". This comparison sometimes leads to generalized assessment that all data contain error, which will be shown not to be true for all data contained in a GIS, but only for the large class of data describing physical reality.

The concept of imperfection has no negative connotation here. On the contrary, it will be argued that imperfections in the data have important beneficial effects. They are necessary to reduce the flood of data reaching us. Imperfections in the data allow us to make decisions with limited cognitive resources. One is reminded of the patient, reported by Sacks (1998), who lost all emotions in an accident that damaged his brain. Common sense could make us believe that a person not disturbed by emotions would be the perfect poker player, become rich with speculations at the stock exchange, etc. Surprisingly, Sacks reported that the man without emotions was unable to lead a normal life because he was *unable to make any decisions*. Emotions seem to be important to pare down the unlimited possibilities for actions to a manageable number. Emotions, which are usually considered disturbing influences in our decision making process, seems to be crucial to our ability to arrive at decisions at all. Imperfections have, as will be argued here, a similar beneficial effect, mostly by reducing the amount of detail.

The article discusses *information processes*, which shall include all processes that

- produce data by *observing* reality,

- transform data, and
- use data to arrive at decision and result in actions.

The term *data* will be used for any information representation, i.e., any physical state (*sign* (Eco 1977)) which is used to represent something else in an information context. Data include observations, represented as values in a computer, as well as text, picture, oral descriptions and mental states. It includes also what is sometimes differentiated as knowledge (albeit an operational, generally accepted definition for knowledge seems not to exist).

All data and knowledge are represented, either in an external medium, e.g., print, computer memory, or in the human cognitive system. Assuming a physical mental representation is not assuming that the format or organization of mental representation is anything similar to the formats used for technical storage, e.g., in a computer. It only assumes that the information humans have in their brains does not exist in a mysterious immaterial form.

The terms situation and context are differentiated in this article. *Situation* describes the circumstances of a cognizant agent in the world, including the environment in which he is situated, as well as his goals, intentions, and needs. *Context* of a sign, e.g., a word, means other signs and how they generally and in the particular instance relate to the sign in focus. Semantic networks (Wikipedia 2007) are an example how the context of aword can be represented (Figure 3).

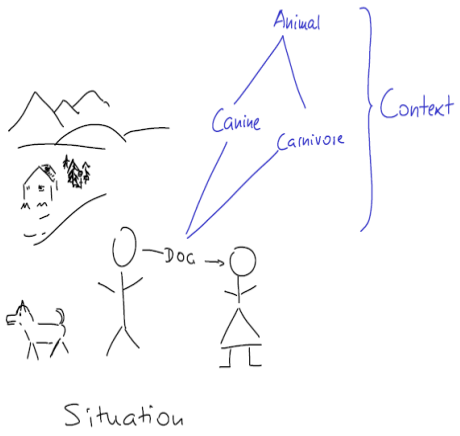


Figure 3: Situation of two agents and context of the utterance “DOG”

# 3 TIERED ONTOLOGY

An ontology describes the conceptualization of the world used in a particular context: two different applications use different conceptualizations. A car navigation system determines the optimal path using the conceptualization of the street network as a graph of edges and nodes, whereas an urban planning application conceptualizes the same space as regions with properties. The ontology clarifies these concepts and communicates the semantics intended by data collectors and data managers, to persons making decisions with the data.

If an ontology for an information system wants to include an assessment of the usability of the data, it must not only conceptualize the objects and processes in reality but must also describe the information processes that link the different conceptualizations and transform between them. This is of particular importance for an ontology that divides conceptualization of reality in tiers (Frank 2001; Smith and Grenon 2004).

Tier O of the ontology is the physical reality, that "what is", independent of human interaction with it. Tier O is the Ontology proper in the philosophical sense (Husserl 1900/01; Heidegger 1927; reprint 1993; Sartre 1943; translated reprint 1993); sometimes Ontology in this sense is capitalized and it is never used in a plural form. In contrast, the ontologies for information systems [Gruber] are written with a lower case o.

## 3.1 Tier 1: Observations

Reality is observable by humans and other cognitive agents (robots, animals). Physical observation mechanisms produce data values from the properties found at a point in space and time.

$v=p(\underline{x}, t)$ [Goodchild]

A value $v$ is the result of an observation process $p$ of physical reality found at point $\underline{x}$ and time $t$. Tier 1 consists of the data resulting from observations at specific locations and times (termed *point observation*); philosophers speak of 'sense data'. In GIS such observations are realized as raster data

resulting from remote sensing [Burrough, Heuvelink, Tomlin], similar to remote sensing and video, our retina performs many such observations in parallel.

### 3.2 Tier 2: Objects

The second tier is a description of the world in terms of physical objects. An object representation is more compact, especially if the subdivision of the world into objects is such that most properties of the objects remain invariant in time (McCarthy and Hayes 1969). For example, most properties of a taxi cab remain the same for hours, days or longer, such as color, size, form; they need not be observed and processed repeatedly. Only location and occupancy of the taxi cab change often and must be regularly observed. The *formation of objects*—what Zadeh calls granulation (Zadeh 2002)—is a complex process of determining the boundaries of objects and then summarizing some properties for the delimited regions. For objects on a table top (Figure 4) a single process of object formation dominates: we form spatially cohesive solids, which move as a single piece: a cup, a saucer, and a spoon.

Figure 4: Simple physical objects on a table top: cup, saucer, spoon

Geographic space does not lead itself to such a single, dominant, subdivision. Watersheds, but also areas above some height above sea level or regions of uniform soil, uniform land management, etc. can be identified (Couclelis and Gottsegen 1997). They are delimited by different properties (Figure 5); geographic objects are mentally classified as suitable for certain interactions, and can therefore overlap.

Figure 5: Fields in a valley: multiple overlap subdivisions in objects are possible.

### 3.3 Tier 3: Constructions

Tier 3 consists of constructs combining and relating physical objects to abstract constructs. These constructs can be socially coordinated or be strictly personal. Constructed reality links a physical object *X* to mean the constructed object *Y* in the context *Z*.

> *"X* counts as *Y* in context *Z"* (Searle 1995, 28)

Constructions relate physical objects or processes to abstract constructs of objects or process type. Constructed objects can alternatively be constructed from other constructed objects, but all constructed objects are eventually grounded in physical objects. (There are no "freestanding *Y* terms" (other opinion: Zaibert and Smith 2004)). The physical object can be a normal object in a situation like the dog in Figure 4 or a sign, which relates to a constructed context (e.g., the written or spoken word "DOG").

## 4 Information Processes

All human knowledge is directly or indirectly the result of observations, transformed in long and complex chains of information processes. The processes sketched in the previous section will be

analyzed in the following sections to understand their effects on data, specifically how they contribute to imperfections in the data.

From the above follows that all imperfections in data must be the result of some aspect of an information process (Figure2). As a consequence, all theory of data quality and error modeling has to be related to empirically justified properties of the information processes. The production of complex theory for managing error in data without empirical grounding in properties of information processes seems to be a futile academic exercise.

## 5 Observations of Physical Properties at Points

The observations of physical properties at a specific point is a physical process that links tier O to tier 1; the realization of which is imperfect in 3 ways

- systematic bias in the transformation of intensity of a property into a (numerical) value,
- unpredictable disturbance in the value produced, and
- observations focus not at a point but over an extended area.

The systematic bias can be included in the model of the sensor and be corrected by a function. The unpredictable disturbance is typically modeled by a probability distribution. For most sensor a normal (Gaussian) probability distribution function (PDF) is an appropriate choice (1).

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

A sensor cannot realize a perfect observation at a perfect point in space or time. Any physical observation integrates a physical process over a region during a time. The time and region over which the integration is performed can be made very small (e.g., a pixel sensor in a camera has a size of 5/1000 mm and integrates (counts) the photons arriving in this region for as little as 1/5000 sec) but it is always of finite size and duration. The size of the area and the duration influences the result.

The necessary finiteness of the sensor introduces an unavoidable scale element in the observations. The sensor can be modeled as a convolution with a Gaussian of the physical reality.

Scale effects are not yet well understood, despite many years of being listed as one of the most important research problems (Abler 1987; NCGIA 1989; Goodchild, Egenhofer et al. 1999). To keep the focus of this article, scale and effects of scale are excluded and not further considered; it is hoped that the conceptual clarification achieved here may contribute later to advancing research in scale effects in information processes. In particular, it should be possible to model the effect of the size of the area observed to the result and the effect of this on object formation and classification.

## 6 Object Formation (Granulation)

Human cognition focuses on objects and object properties. We are not aware that our eyes (but also other sensors in and at the surface of our body) report point observations, e.g., the individual sensors in the eye's retina give a pixel-like observation, but the eye seems to report about size, color, and location of objects around us. The observations are, immediately and without the person being conscious about the processes involved, converted to object data. These processes connects tier 1 to tier 2. Such processes are found not only in humans; higher animals also form mental representations of objects[]. Object formation increases the imperfection of data—instead of having detailed knowledge about each individual pixel only a *summary description* (summary value) of, for example, the wheat field in Figure 5 is retained, containing color value, size etc.. The very substantial reduction in size of the data is achieved with an increase in imperfection. For example in Figure 4, the area for the big field includes approx. 1.5 M pixels each of which has 8 bits per color channel, for a total of 4.5M bytes. The compact representation as a region requires few points for the boundary, each using twice 4 bytes to represent coordinates and a few bytes to describe the average color of the region (say 10 points in the boundary, for a total of less than 100 bytes). Even if this computation assumes computer technology and is not representative for processes in a human brain, it gives a general idea of the $1:10^5$ compression achieved.

Object formation consists of two information processes

- boundary identification

- computing summary descriptions,

which will be sketched in the following two subsections (more details in [Frank …]), before mental classification is addressed in the following section.

### 6.1 Boundary Identification

Objects are—generally speaking—regions in 2D or 3D that are uniform in some aspect. The field in Figure 5 is uniform in its color, tabletop objects in Figure 4 are uniform in the material coherence and in their movement: each point of the rigid object moves with a corresponding movement vector. Note that object formation exploits the strong correlation found in the real world; human life, would not be possible in a world without strong spatial and temporal correlation [Goodchild]. The details of how objects are identified is determined by the interactions intended and the situation. The focus of this article excludes a detailed discussion of processes in reality and how they interact with objects but processes depend on properties of the object involved—thus determining their boundaries. Processes can be granulated by similar approaches in a 3D plus time space.

An object boundary is determined by first selecting a property and a property value that should be uniform across the object. This process is similar to the well-known procedure for regionalization of 2D images []. It produces a region of uniform values and boundaries for these regions. This process is useful to identify the 2 D regions in an image. These uniform regions are candidates for being projections of 3D objects of interest (Marr 1982). Two different methods for determining object boundaries have been identified (perhaps others are possible):

A) By *thresholds* on the values of interest: the object is the connected region of all point observations for which the value $v$ is between two limits $v_L$ and $v_H$
B) By *maximal change*: the object boundary is where the value $v$ changes maximally (Burrough 1996).

The location of the boundary derived by these two methods is not the same and the applicability of the two methods differs: Method A applies where a condition must be satisfied for an object (e.g.,

minimal annual rainfall to determine where wheat can grow) and method B applies where separation of object from environment is important, (e.g., to delimit a mountain). Additionally, object formation may use rules that require minimal size for an object; these rules cause a scale effect in the object forming process and will not be discussed here further.

Assuming a PDF for the determination of the boundary one can describe the PDF for the boundary line from each of the two methods discussed above. The information process has an associated transformation function that transforms the PDF of the point observation in a PDF for the boundary line (Figure 6).

### 6.2 Determination of Descriptive Summary Data

Descriptive values, summarize the properties of the object determined by a boundary. The computation is typically an integral or similar summary function that determines the sum, maximum, minimum, or average over the region, e.g., total weight of a movable object, amount of rainfall on a watershed, maximum height in country (Tomlin 1983; Egenhofer and Frank 1986).

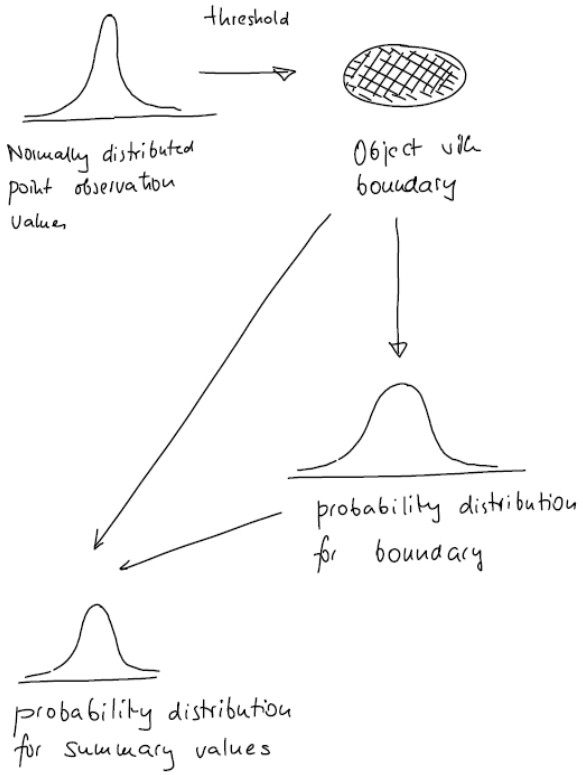


Figure 6: Transformation of probability, distinction functions from observations to boundary and summary value

Given the PDF for the value of interest of the summary (which is not necessarily the same value with the same PDF as the property values used to determine the boundary) and the PDF for the boundary, a PDF for the summary values is obtained by transformation ofthe input PDF (Figure 6).

If the observation information processes allow a probabilistic description of the imperfections of the values, than the imperfections in the object boundary and summary value are equally describable by a probability distribution. It is an interesting question to determine if the PDF transformation functions associated with boundary derivation and derivation of summary values preserve a normal distribution, i.e., if the observation processes described by imperfections with a normal distribution, produce imperfections in boundary location and summary values which are describable by a normal distribution.

## 7 Classification

Objects once identified are classified. On the tabletop, we see glasses, forks, and plates; in a landscape forest, fields, and lakes are identified; mental classification is an information process within to tier 2. It is well known that classification of objects by humans is a complex and multifaceted process; here I only address the cognitive (subjective, personal) mental classification of objects with respect to a potential interaction, not the taxonomy fixed in a vocabulary. Gibson has used the term affordance for the potential use of an object (Gibson 1986; Raubal 2002). I develop here a generalization for a formalizable theory of affordances from mental classification (Raubal 2002). It separates mental process from linguistic approaches typically for classification research; linguistic classification produces most of the vocabulary of natural languages where: nouns, like ‘dog’, or ‘forest’ describe classes of objects, not individuals. The effects of language related classifications will be studied later (subsection 8.4)

Figure 7: Pouring requires two container objects and one liquid object

Mental classification relates the objects identified by granulation processes to operations, i.e., interactions of the cognitive agent with the world. To perform an action, e.g., to pour water from a pitcher into a glass (photo Figure 7) requires a number of properties of the objects involved: the pitcher and the glass must be containers, i.e., having the affordance to contain a liquid, the object poured must be a liquid, etc. Empirical evidence [...] shows that mirror neurons (Rizzolati, Craighero et al. 2002) found in humans and (at least) apes classify not only operations the cognizant agent sees (i.e., visually perceives) but also classify the objects with respect to having the right properties to be involved in an operation. Potential interactions between the agent and objects (or interactions of interest between objects) require conditions these objects must fulfill, expressed as a property and a range for the value of the property. In this way operations of interest indicate what properties are important and these important properties are then used to determine boundaries of objects (above subsection 6.1) (Frank 2006)[Kuhn].

I have used the term distinction for the differentiation between objects that fulfill a condition and those that do not (Frank 2006). Distinctions are partially ordered: a distinction can be finer than another one (e.g., drinkable is a subtaxon of liquid), distinctions form a taxonomic lattice (Frank 2006). The mental taxonomy adapts in the level of detail to the situation and can be much finer if the situation requires it than the one implied in the vocabulary (Ganter, Stumme et al. 2005).

Humans classify unconsciously and immediately the objects we encounter and retain only the classification (without applying verbal labels). Grouping of distinctions required for typical

interactions form abbreviations for sets of often required properties of objects. For example: the flat things that can be cut by a pair of scissors (i.e., paper), or the self-powered, movable things steered by a human passenger (i.e., cars). The classification in the mental taxonomic lattice is an abstraction reducing the amount of detailed information initially perceived in preparation for a probable decision. Instead of retaining detailed values for the decisive properties till the time of decision making only the classification is retained. This is likely what Gibson meant when he stressed that classification by affordances is without representation (Gibson 1986).

This abstraction process is cognitively plausible and supported by empirical evidence. If you interact with a household object (e.g., eat from a plate in a restaurant) and are later asked about detailed properties of the object you most likely realize that the properties you considered to classify the object as a plate were not retained, only the final classification. The situation influences the interactions with the objects an agent considers; the relevant interaction determines which properties to use for object formation. All of this is summarized in a classification.

Distinctions reflect the limits in the property values of an object, where the object can or cannot be used for a specific interaction. The decision whether the values for an object are inside the limits or not is more or less sharp and the cutoff gradual (Figure 8). The distinctions and classifications are therefore fuzzy values, i.e., membership functions as originally defined by Zadeh (1974).

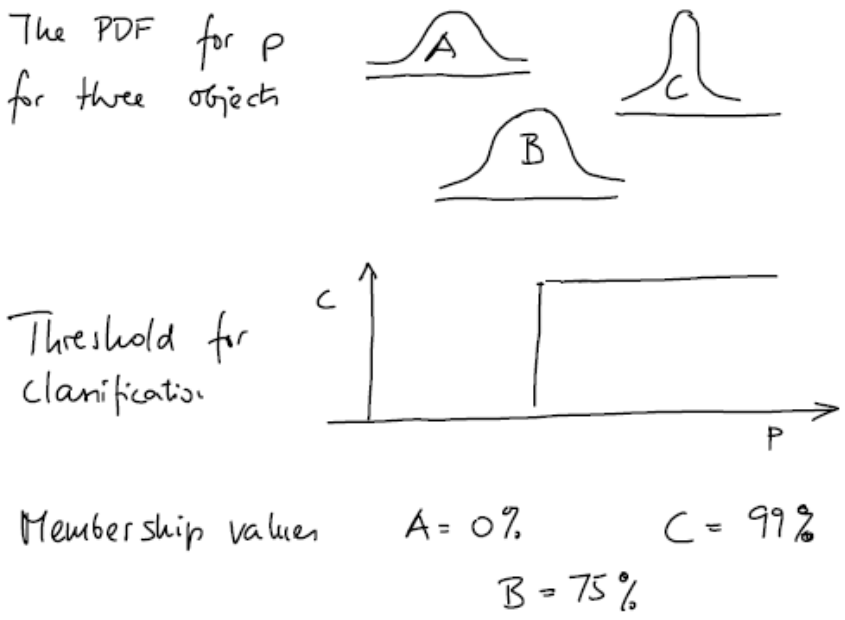


Figure 8: Classification of objects result in fuzzy membership values

## 8 CONSTRUCTIONS

In tier 3 the world of *constructions* is discussed, which is linked through granulation and mental classification to the physical reality of physical objects and operations. By construction I mean concepts that are (1) mental units, which (2) have external representations (signs, e.g., words), (3) can be communicated between cognitive agents, and (4) are, within a context, without imperfection or error.

In addition to the agent's direct sensory experience of the world and as a reflection of the agent's experience of the world, an externally representable information image of reality is created duplicating the sensory "reality" in the brain. I call the constructions that stand for direct experiential reality *grounding items*. The duplication of experience and the construction is mostly an artifact necessary for description of the present theory; the sensory experience and the grounding items are isomorphic and are not consciously separable. This representable image of the world is here for explanatory purpose separated from the (subjective) experiential concepts in the cognitive agent's mind (Figure 9).


Figure 9: The grounding of constructs in experiential concepts

The representable images are constructed as models of reality. These models may be a verbal description, oral or written, a computational model, a sketch, etc. and strongly interconnected by operations and relations. I describe such models as algebras and posit that they are—in a fuzzy way [Zadeh London 2007]—homomorphic to reality [Kuhn, Frank…].

### 8.1 Context

The meaning of constructions are determined in a web of concepts that are bound by the relations between the concepts (semantic web). The full set of concepts that are interrelated are called a context; the semantics of the concepts are determined only through the relations in this context and within this context. The meaning of a construction is in the structure and requires a context.

Considering these structures as algebraic structures indicates that the semantics is determined only up to a structure preserving isomorphism. This is not a limitation and an uncertainty but is the precondition for communication to be possible: it must be possible to translate between different representations (mental, verbal, written). To maintain the meaning; the translations must be structure preserving mappings (Eco 2003).

### 8.2 Grounding

Using Searle's formula for a semi-formal treatment, I posit that mental experiential concepts have corresponding representable concepts, both for individuals and for classes.

*"X* counts as *Y* in context *Z"*

can be generalized: an experiential concept counts as a representable concept in a context. Note that the experiential concept—an experience of a thing in reality—can be caused by an ordinary physical object or a physical object that is intended as a sign (Eco 1976) this justifies the generalization of Searle's formula. The formula provides grounding for all constructions in mental concepts, which are all directly or indirectly related to such grounding items and through these experientially grounded.

### 8.3 Communication

Despite the fact that we do not know exactly how humans learn their mother tongue [...] it is an empirical observation that humans establish a consensus on the meaning of external signs; human communication is possible, even though it is not perfect. Acquiring a language means to establish a correspondence between experiential concepts and constructions. The "fuzzy homomorphism" between experience and mental models which must be reflected in the verbal communication seems to be sufficient to converge into a common encoding over repeated experiences. The fact that initial language acquisition occurs in a simplified reality and within a supportive affective environment may significantly influence how the mechanism of language acquisition works.

### 8.4 Imperfections in Communication

The meaning of a sign is defined in its context and this context can vary between sender and recipient of a sign. If a sign is unique to a context, no confusion is possible, but for homonyms (same sign in different context) a potential for imperfect communication exists. Natural language is rich in polysemy, where the same word (sign) means in different contexts different things. WorldNet (Fellbaum 1998) documents the polysemy by separating different meanings of a word in synsets.

The imperfection of communication increases with the distance between the contexts. A description of a soil for civil engineering, hydrology, or agriculture may use the same words, but the meaning is different because the words are in each science connected differently; the structure established in each context is different. International organizations, e.g., ISO, establish formalized contexts in which signs can be exchanged with the assumption of unchanged semantics for particular application areas (e.g. hydrology, commerce).

In normal communication circumstances, multiple contexts are combined. For example, participants in a meeting each have a subjective component in their context as well as a role influenced context, part of discussion in meetings serves to align the contexts of the participants (Rottenbacher 2006).

Language classifies (constructions of) objects; this classification is very similar and reflects similar concerns as mental classification (section 7 above). Separating mental and linguistic classification is a step to classify the ontology of information process. Previous research in the semantics of linguistic classification has identified a radial structure (radial categories (Rosch 1973)). The same code in multiple contexts has widely overlapping applicabilities, which share a core meaning (Figure 10).

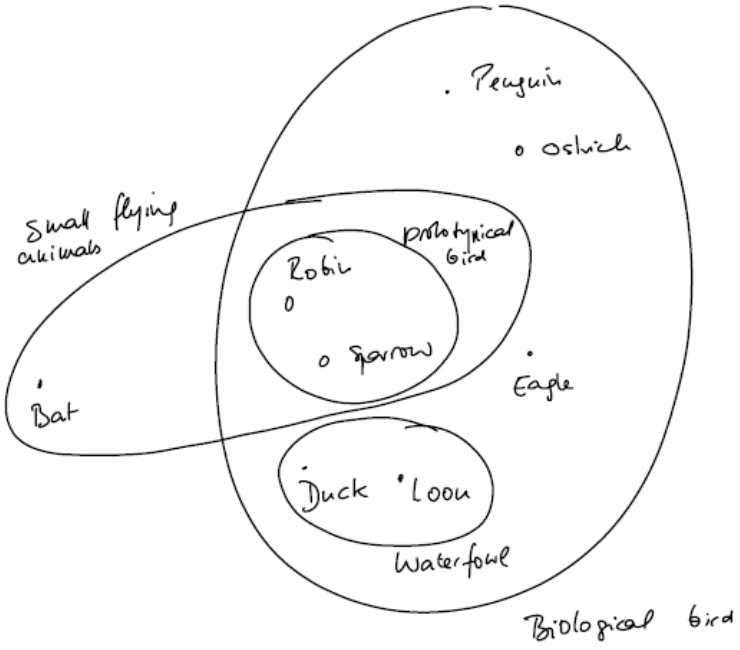


Figure 10: Different meaning of 'Bird' in different contexts

Exemplars in the area of overlap are coded in each context with the same code. Other exemplars are coded differently, depending on the context. One speaks of "better examples" for a class (e.g., robin of sparrow are "better" birds than penguin or ostrich), which is in contrast to a set theoretic approach, where an exemplar is or is not member of the class and no gradual membership is possible.

The theory of supervaluation [] gives guidelines how to deal with the integration of multiple context and reasoning in an integrated data collection. If the semantics of the context are available as formally described ontologies (e.g., described in OWL (Dieckmann 2003)) then formal conversions between codes from different contexts can be attempted and the effects assessed or compensated.

A probabilistic approach, for example considering overlap area to total area cannot lead to a solution, because differences in context that result in different taxonomy are correlated with the purpose of the construction of the taxonomy and influenced by the situation. For example, the classification "snake" does not justify a 50% probability that the snake in the garden is dangerous. It depends on the geographic region whether venomous snakes occur or not, and 'snake' may imply 'harmless snake' or 'dangerous snake' depending on previous experience in the geographic context.

The translation between contexts must always relate the signs in the context back to the grounding items and then forward to the sign of interest in the other context (in Figure 11 the grounding items are shown duplicated, as experiential mental concepts and grounding items, which are isomorphic and are separated here only for clarity of the figure).

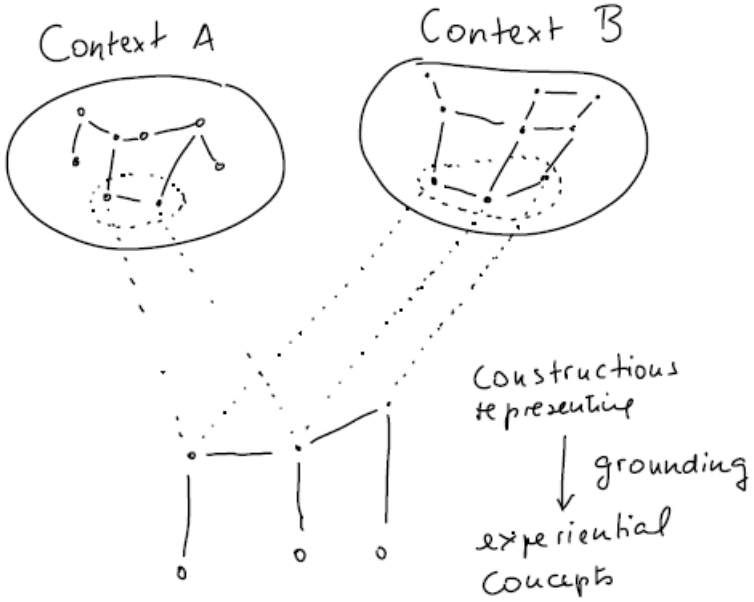


Figure 11: Translating between contexts in one individual through experientially grounded concepts

The rich structure of reality as we experience it is approximately shared by all humans (Lakoff 1987), because they share a large part of daily experiences (eating, drinking, sleeping, ...). This approximation of the experiential grounding is usually sufficient to establish a mapping between structure encountered in a text received and our own structure among constructs (Figure 11 cannot show the density of the relation that determines the structure and which force a mostly correct mapping).

### 8.5 Constructions Are without Imperfections

Constructions are, unlike observations (in tier1) descriptions or classification of objects (of tier2) without imperfection and error, as long as they are used in a fixed and shared context. As an everyday example, consider a description of the paper bill in (Figure 12):


Figure 12: Paper and metal objects, counting as money in the Czech legal context

The length of the paper is observed as 134 mm, with a standard deviation of 3/100 mm, but once established that this is a Czech bill of "Padesát Korun Českých" there is no uncertainty in the 50 Czech Crowns value; it is not 49.90 or 50.05 with a probability! In the context of Czech commerce, it is 50 without error or imperfection. If we leave this context, then the value expressed in Euros may be uncertain, today the exchange rate is 33.2050 Crowns per Euro, which gives an approximate value of 1.50 Euro for the 50 Crown bill.

The value 50 is here—unlike the measured physical length—without error, directly contradicting the often heard statement that "all data contains some error", which is correct only for tier 1 and 2, but not for the constructions of tier 3. Mathematics is the best example of a construction that is without error (except for human failure to work correctly). The difficulty and imperfections of tier 3 are introduced by

- establishing the connection between experiential concepts and constructions—the 'subsumption' of the legal profession where one establishes whether a concrete act was 'murder' or 'manslaughter', and
- translating between contexts.

## 9 Differences to Previous Descriptions of Tiered Ontology

I have initially presented the concept of a tiered ontology in a workshop in 2000 (Frank 2001) and later extended the description for spatio-temporal data (Frank 2003). A number of recent articles have explored the processes that produce and transform information up to the tier 2 of physical objects and analyzed how such processes influence data quality [see drafts on my web page]. The third tier (or tiers 3 and 4) including classification as part of language and was originally separated into social construction and subjective data. This division was not satisfactory and later publications were less specific. The present focus on information processes separates mental and linguistic classification and includes in tier 3 all constructions. I posit that Searle's description of social construction can be generalized.

All parts of tier 3 (i.e., not the description and mental classification of physical objects) are constructions and depend on the context in which the semantics of the construction is defined. This hypothesis could be falsified by a single example of a higher level concept that cannot be shown to be a context dependent construction and externally representable. I have not found one yet!

For example, mental states and emotions are not constructions but proprio-sensor related observations of states of the agent; as a consequence we cannot communicate them directly! The translation to a (e.g., verbal) sign allows imperfect communication limited by the imperfect indirect knowledge of internal states of others obtained through observations of secondary signals (e.g., body language). The translation between the contexts of emotion terms between two persons is quite imperfect. Mathematics as an extreme other example, is all construction and meanings of communicatable signs are defined only through the context, without experiential grounding. Mathematics is motivated by observation in the world but is in its abstract form the cleanest example of a construction (Lakoff and Nuñez 2000). The legal system is similar in construction. When the law is applied to render a judgment, subsumptions are used to link between abstract notions in legal text (e.g., man slaughter) and real objects and processes (e.g., the events in the dark night of March 26, 2002 that caused the death of a person).

# 10 Further Research

The framework presented suggests a number of specific research topics.

## 10.1 Granulation

Assume two fields of point observations of two properties *a*, *b* with a normal distribution and with standard deviations $\sigma_A$ and $\sigma_B$. What is the probability distribution function (PDF) for a summary value *B*, which is the sum (integral) of *b* inside the boundary? Example case to work out: determine the boundary of a mountain (> 1000m) and determine the area [Navratil paper].

## 10.2 Classification

To work out an example of the relation between object formation, mental classification and observed properties would be useful and with an algebraic specification language feasible. Can the proposed process of thresholding for classification lead to numeric assessment of the fuzzy membership values? (As suggested in Figure 8).

## 10.3 Constructions

Supervaluation is a very coarse approach to the semantics of the same words used in different contexts. An analysis of two contexts for the same term should indicate a path towards a transformation of constructions between contexts. Of particular interest is the question, how much common grounding between the contexts is required (Figure 10). It could be useful to observe how the banking industry deals with changing money between national contexts and to study the rules established in international law to convert concepts between legal contexts.

## 10.4 Scale Effects

The consistent framework for the processing of information across various levels provides a backdrop to investigate scale effects and to formalize their effects. The Modifiable Areal Unit Problem for observation can be formalized: given a fixed field of values for a property *p*, which is observed with an imperfect non-point observation instrument. The observation can be modeled by

convolutions that simulate the instant field of vision of the observation instrument. The effect of scale introduced by the observation instrument and other effects along the information processing chain can be traced and the effects on the decision assessed.

## 11 Conclusions

If one assumes that all knowledge we have must come from observations—where else should knowledge come from?—then the processes of transformation of observations to knowledge comes into focus. The imperfection in our knowledge must be caused by imperfections in these processes.


Figure 13: Imperfections in knowledge must be produced by imperfections in information processing

This article combines a topic of information system ontology and uses a tiered ontology to structure all imperfections in our knowledge. As a result, the third ontological tier is restructured as generalized constructions that are described by Searle's rule

*X* counts as *Y* in the context *Z*, (Searle 1995, 28),

which is here used not only for what Searle called social construction but generalized. The (new, reduced) tiered ontology structures the knowledge of the world in 3 tiers

- Observations of properties at a point in space and time
- Attributes and classification of physical objects derived by granulation from observations and their mental classification
- Constructions of knowledge in symbolic systems.

The observation processes are imperfect physical processes and are influenced by random distributed errors and scale effects. The analysis of the information processes shows that imperfections in the observations are, in first approximation and as a very general conclusion, randomly distributed. The granulation processes form objects with uncertain boundaries and summary attribute values. The imperfections in object boundaries and summary attribute values result from a transformation of the random observation errors and can be modeled by probability functions. The imperfections can be derived as transformations from the errors in the observations.

Mental classification compares the observed attributes of objects with the requirements to interact with them (or for interactions among them). The classification uses thresholds of the attributes values that determine whether an object can interact with others in a specific way; imperfections are likely best modeled by the theory of fuzzy values.

The constructions $Y$ are grounded in experiential objects $X$ and externalize the mental classifications. Constructions are represented as signs (words, codes) and the semantics is in the (algebraic) structure given by the context, which permits transformation between representations. The constructions can be very precise within a single context, but most interesting reasoning combines knowledge across contexts. The constructions and the related symbolic processes are in principle perfect—as perfect as logical reasoning in mathematics—*within a single context*. Actual processes of communication and production of higher level of knowledge are always combining knowledge from different contexts, e.g., by different people involved with observations and previous steps in symbolic information processing. The imperfections arise from the incompatibilities between the contexts and an approach with supervaluation appears promising.

The study of real world processes (Frank draft 2007) should strengthen the arguments made here. The situation of a cognitive agent, i.e., his environment and his goals, determine which operations and properties of objects are relevant. The required properties of objects are then selected to determine the boundaries of objects, summary values relevant and produce situation-oriented classifications. Communication of such situation-oriented classification is facilitated if the other agent can see the situation.

The analysis has intentionally not included scale effects and real processes. I hope that the framework established here will lead to fruitful investigations of these very important questions in the future.

As a final word, I want to stress the importance and positive contributions of imperfections in the information process: imperfections serve to reduce the immense amount of possible data that result from the observation of the world. Together with the reduction in detail comes necessarily a reduction in the perfection of the knowledge. The solution is not to try to produce perfect knowledge —the metaphorical map in 1:1 scale [Borges, Carol]—but to manage the imperfections to achieve maximal compression by minimal errors in decisions. Imperfections in knowledge are mostly beneficial!

## Acknowledgements

My daughters Stella and Astrid have taught me much about how we learn about the world. To bring this into a scientific context and to sharpen the analysis to make it presentable was a long process; I am in debt to many: my former colleagues from the NCGIA, especially David Mark, Mike Goodchild, Max Egenhofer, Werner Kuhn and former students and colleagues in Vienna, Martin Raubal, Damir Medak, Gerhard Navratil, Christine Rottenbacher and Florian Twaroch. Discussions with Barry Smith, John Searle and other philosophers have helped me progress to this current state of ignorance.

## REFERENCES


Abler, R. (1987). "The National Science Foundation - National Center for Geographic Information and Analysis." IJGIS **1**(4): 303-326.

Burrough, P. A. (1996). Natural Objects with Indeterminate Boundaries. Geographic Objects with Indeterminate Boundaries. P. A. Burrough and A. U. Frank. London, Taylor and Francis**:** 3-28.

Chrisman, N. R. (1989). Modeling Error in Overlaid Categorical Maps. The Accuracy of Spatial Databases. M. Goodchild and S. Gopal, Taylor & Francis**:** 21-34.

Couclelis, H. and J. Gottsegen (1997). What Maps Mean to People: Denotation, Connotation, and Geographic Visualization in Land-Use Debates. Spatial Information Theory - A Theoretical Basis for GIS (International Conference COSIT'97). S. C. Hirtle and A. U. Frank. Berlin-Heidelberg, Springer-Verlag. **1329:** 151-162.

Dieckmann, J. (2003). DAML+OIL und OWL XML-Sprachen für Ontologien. Berlin**:** 21.

Eco, U. (1976). A Theory of Semiotics. Bloomington, Indiana University Press.

Eco, U. (1977). Zeichen - Einfuehrung in einen Begriff und seine Geschichte. Frankfurt a. Main, Edition Suhrkamp.

Eco, U. (1993). Die Suche nach der vollkommenen Sprache. Muenchen, Deutscher Taschenbuch Verlag.

Eco, U. (2003). Dire quasi la stessa cosa. Milano, Bombiani.

Egenhofer, M. and A. U. Frank (1986). Connection between Local and Regional: Additional "Intelligence" Needed. FIG XVIII International Congress of Surveyors, Toronto, Canada (June 1-11, 1986).

Fellbaum, C., Ed. (1998). WordNet: An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, Mass., The MIT Press.

Frank, A. U. (2001). "Tiers of ontology and consistency constraints in geographic information systems." International Journal of Geographical Information Science **15**(7): 667-678.

Frank, A. U. (2001). "Tiers of Ontology and Consistency Constraints in Geographic Information Systems." International Journal of Geographical Information Science **75**(5 (Special Issue on Ontology of Geographic Information)): 667-678.

Frank, A. U. (2003). Ontology for Spatio-Temporal Databases. Spatiotemporal Databases: The Chorochronos Approach. M. Koubarakis, T. Sellis and e. al. Berlin, Springer-Verlag. **2520:** 9-78.

Frank, A. U. (2006). Distinctions Produce a Taxonomic Lattice: Are These the Units of Mentalese? International Conference on Formal Ontology in Information Systems, Baltimore, Maryland, IOS Press.

Frank, A. U. (draft 2007). Ontologies for Imperfect Data in GIS, Department for Geoinformation and Cartography, TU Wien**:** 22.

Ganter, B., G. Stumme, et al., Eds. (2005). Formal Concept Analysis Foundations and Applications. Berlin, Heidelberg, Springer.

Gibson, J. J. (1986). The Ecological Approach to Visual Perception. Hillsdale, NJ, Lawrence Erlbaum.

Goodchild, M. F., M. J. Egenhofer, et al. (1999). "Introduction to the Varenius Project." International Journal of Geographical Information Science **13**(8): 731-745.

Heidegger, M. (1927; reprint 1993). Sein und Zeit. Tübingen, Niemeyer.

Husserl (1900/01). Logische Untersuchungen. Halle, M. Niemeyer.

Lakoff, G. (1987). Women, Fire, and Dangerous Things: What Categories Reveal About the Mind. Chicago, IL, University of Chicago Press.

Lakoff, G. and R. E. Nuñez (2000). Where Mathematics Comes From - How the Embodied mind Brings Mathematics into Being, Basic Books.

Marr, D. (1982). Vision. New York, NY, W.H. Freeman.

McCarthy, J. and P. J. Hayes (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. Machine Intelligence 4. B. Meltzer and D. Michie. Edinburgh, Edinburgh University Press**:** 463-502.

NCGIA (1989). "The Research Plan of the National Center for Geographic Information and Analysis." International Journal of Geographical Information Systems**3**(2): 117 - 136.

NCGIA (1989). "The U.S. National Center for Geographic Information and Analysis: An Overview of the Agenda for Research and Education." IJGIS **2**(3): 117-136.

Openshaw, S., M. Charlton, et al. (1987). "A Mark 1 Geographical Analysis Machine for the Automated Analysis of Point Data Sets." International Journal of Geographical Information Systems **1**(4): 335-358.

Pickles, J., Ed. (1995). Ground Truth - The Social Implications of Geographic Information Systems. Mappings: Society/Theory/Space. New York, London, The Guilford Press.

Raubal, M. (2002). Wayfinding in Built Environments: The Case of Airports. Münster, Solingen, Institut für Geoinformatik, Institut für Geoinformation.

Rizzolati, G., L. Craighero, et al. (2002). The Mirror System in Humans. Mirror Neurons and the Evolution of Brain and Language. M. Stamenov and V. Gallese, John Benjamins Publishing Company**:** 37-59.

Rosch, E. (1973). "Natural Categories." Cognitive Psychology **4**: 328-350.

Rottenbacher, C. (2006). Bewegter Planungsprozess. Department of Geoinformation and Cartography. Vienna, Technical University Vienna. **PhD**.

Sacks, O. (1998). The Man who Mistook his Wife for a Hat. New York, Touchstone (Simon & Schuster).

Sartre, J. P. (1943; translated reprint 1993). Being And Nothingness. New York, Washington Square Press.

Searle, J. R., Ed. (1995). The Construction of Social Reality. New York, The Free Press.

Smith, B. and P. Grenon (2004). "SNAP and SPAN: Towards Dynamic Spatial Ontology." Spatial Cognition and Computing(4): 69-103.

Tomlin, C. D. (1983). A Map Algebra. Harvard Computer Graphics Conference, Cambridge, Mass.

Wikipedia. (2007). "Semantic Networks." from http://en.wikipedia.org/wiki/Semantic_network.

Zadeh, L. A. (1974). "Fuzzy Logic and Its Application to Approximate Reasoning." Information Processing.

Zadeh, L. A. (2002). Some Reflections on Information Granulation and Its Centrality in Granular Computing, Computing with Words, the Computational Theory of Perceptions and Precisiated Natural Language. Data Mining, Rough Sets and Granular Computing. T. Y. Lin, Y. Y. Yao and L. A. Zadeh. Heidelberg, Germany, Physica-Verlag GmbH**:** 3-20.

Zaibert, L. and B. Smith (2004). Real Estate - Foundations of the Ontology of Property. The Ontology and Modelling of Real Estate Transactions: European Jurisdictions. H. Stuckenschmidt, E. Stubkjaer and C. Schlieder, Ashgate Pub Ltd**:** 35-51.