


Stefan Schulz

Medical Informatics Research Group

University Medical Center

Freiburg, Germany

Named Entity or Entity Name?

Elucidating a Dubious but Popular Concept

Dagstuhl Seminar 08131 - Ontologies and Text Mining for Life Sciences: Current Status and Future Perspectives

UNIVERSITÄTS KLINIKUM FREIBURG



Named Entities

- Named Entity Recognition (NER): identification of non-lexicalized portions in free text.
- “Named Entity” or, simply, “Entity”: Everybody uses this term, but what does it really mean ?
- First used in MUC-6 (Message Understanding Conference) “named entities task”: “*entity names were defined as proper names, acronyms and other unique identifiers which can be categorized in terms of organizations, persons and locations*”
- MUC-7: “named entity”: “*proper names and quantities of interest*”

Two readings of “Named Entity”

- NE_1: unit of language, i.e., a linguistic sign, commonly made up of a string of alphanumeric and punctuation characters denoting some concrete or abstract entity in the real world;
- NE_2: entity in the real world (outside the realm of language) which is characterized by being referred to by a linguistic sign that is considered a “name”

Credentials for NE_1 (language objects)

Named entities are…

- "…atomic units, such as proper names, temporal expressions (e.g., dates) and quantities" [Bojar O, Homola P, Kubon V: An MT System Recycled. In Proceedings of MT Summit X 2005:380–387]
- „…phrases that contain the names of persons, organizations, locations, times and quantities" [Sang E: Introduction to the CoNLL-2002 Shared Task: Language-independent named entity recognition. In CoNLL-2002 – Proceedings of the 6th Conference on Natural Language Learning.]
- "…terms referring to individuals, locations organizations, and dates" [Tonella P, Ricca F, Pianta E, Girardi C: Using Keyword Extraction for Web Site Clustering. wse 2003, 00:41]

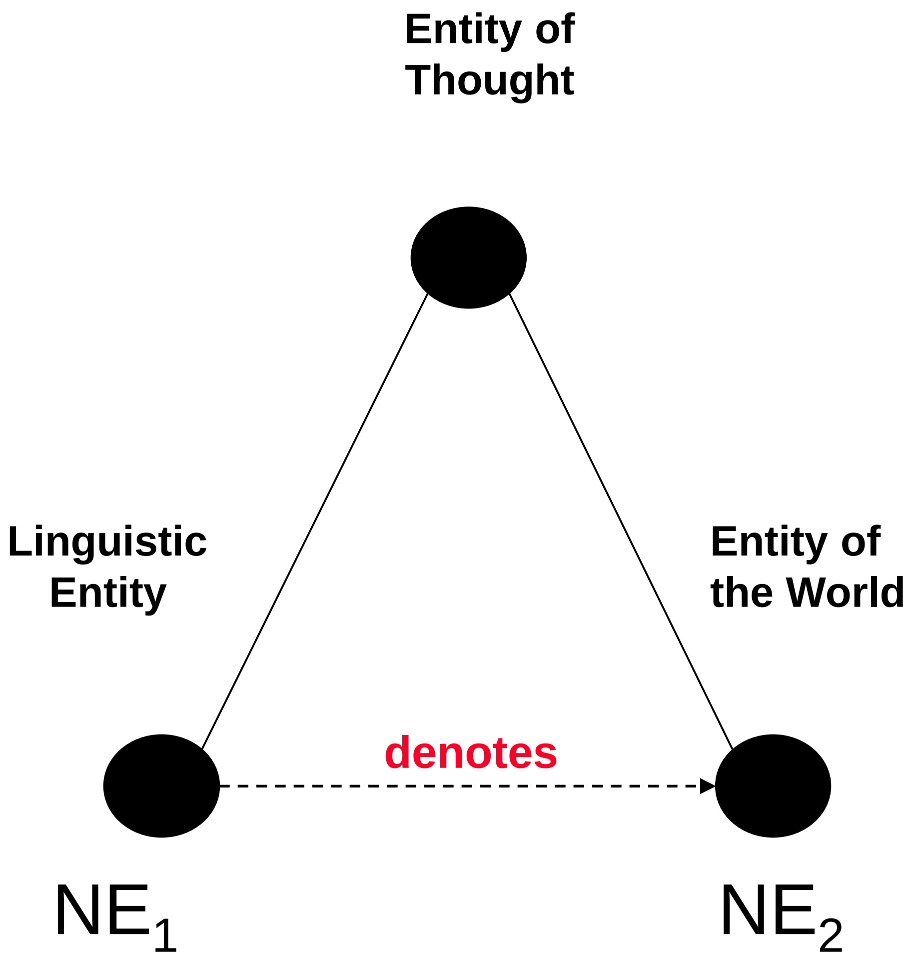
Credentials for NE_2 (domain objects)

- "Diversified named-entities are involved in different sorts of interactions in the specific domain"
  [Song Y, Kim E, Lee GG, Yi BK: POSBIOTM—NER: a trainable biomedical named-entity recognition system. Bioinformatics 2005, 21(11):2794–2796.
- "A named entity is a person, place, etc. that has a specific name. For instance, Microsoft and George Bush are both named entities"
  [Brown J: Entity-Tagged Language Models for Question Classification in a QA System. Tech. rep., Carnegie Mellon University, http://www.cs.cmu.edu/jonbrown/IRLab/Brown-IRLab.pdf 2006.]
- "names of named entities"
  [Kim JD, Ohta T, Tatisi Y, Tsujii J: GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics 2003, 19:180–182.]

More confusion

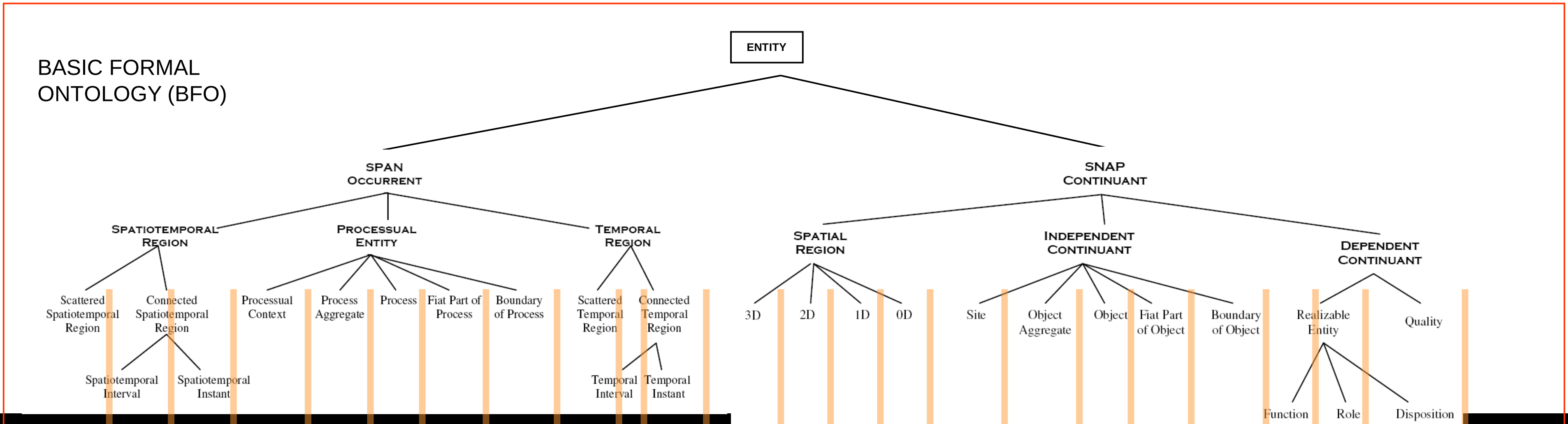
- “Named entities are people, organizations, locations and others referred by name. The wide interpretation of the term includes any tokens referring to something specific in the world: numbers, addresses, amounts of money, etc.” [Popov B, Kiryakov A, Kirilov A, Manov D, Ognyanoff D, Goranov M: KIM - Semantic Annotation Platform. In International Semantic Web Conference, LNCS 2870]
- “…recover “biological entities from free text” [Krallinger M, Valencia A: Text-mining and information-retrieval services for molecular biology. Genome Biology 2005, 6(7):224.]
- “Currently, text-mining applications are being employed in the identification of biological entities, such as protein or gene names” [Krallinger M, Alonso-Allende R, Valencia A: Text-mining approaches in molecular biology and biomedicine. Drug Discovery Today 2005, 10:439–445.]

Principal source of confusion



Entity

- Latin *ens* = being
- In CS: element in a data structure: *“entity relationship model”*
- In Philosophy: something that has separate and distinct existence in an objective or conceptual reality (are there non-entities ?).
- In formal ontologies:



The term „Entity“ encompasses

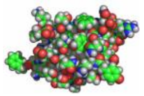
- Intralinguistic Entities:
  - Elements of language, e.g. words, terms, phrases
- Extralinguistic Entities:
  - Particular objects of the real world (People in this room, a DNA strand in a cell, a concrete transcription process)
  - Types (classes) of objects (Human, DNA, transcription process) that are instantiated by particulars
  - Human thoughts, concepts, plans (blueprint of a molecule, right femur bone in an atlas of anatomy, *bauplan* of Airbus A 380, manned Mars mission…)

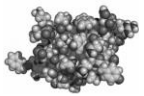
Term vs. Name

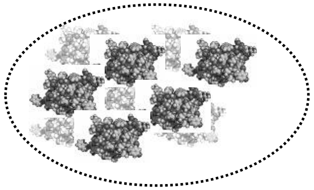
- ISO TC37
  - Term: “designation of a defined concept in a special language by a linguistic expression”
  - Name: “designation of an object by a linguistic expression”
- Observation: medicine -> “term”
  biology -> “name”
- Can we map the name / type distinction to the particular / type distinction ?
- Example:
  - {Clyde$_{\text{NAME}}$} is an instance of {Asian elephant$_{\text{TERM}}$ }
  - {Schloss Dagstuhl$_{\text{NAME}}$} is an instance of {Building$_{\text{TERM}}$ }

More complicated with biomedical names

- "Human insulin": Name or term ?
- Ambiguity due to concurrent readings:
  1. A blueprint of an insulin molecule (particular)
  2. Insulin Molecule (type)
  3. Collection of Insulin molecules (type)

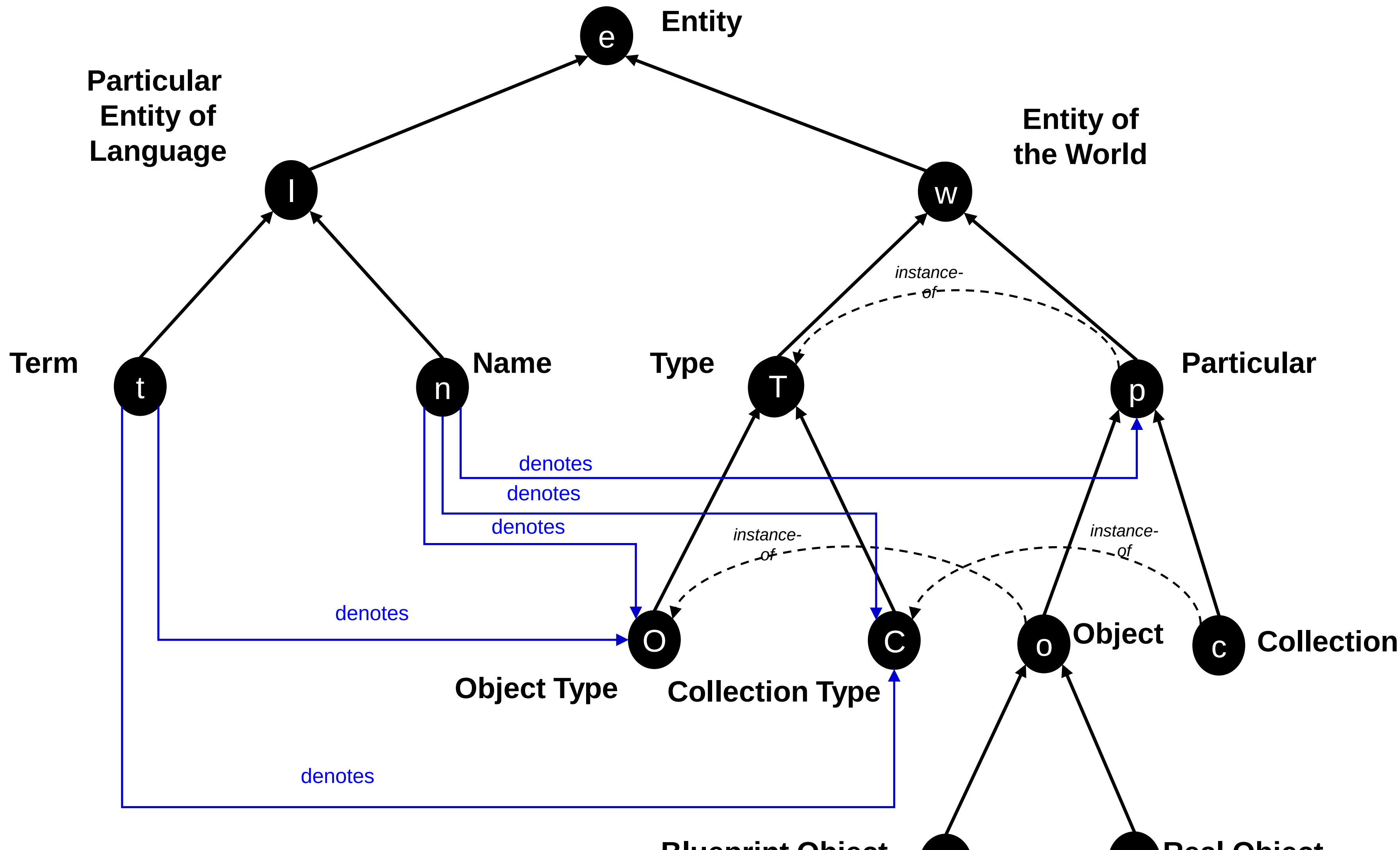


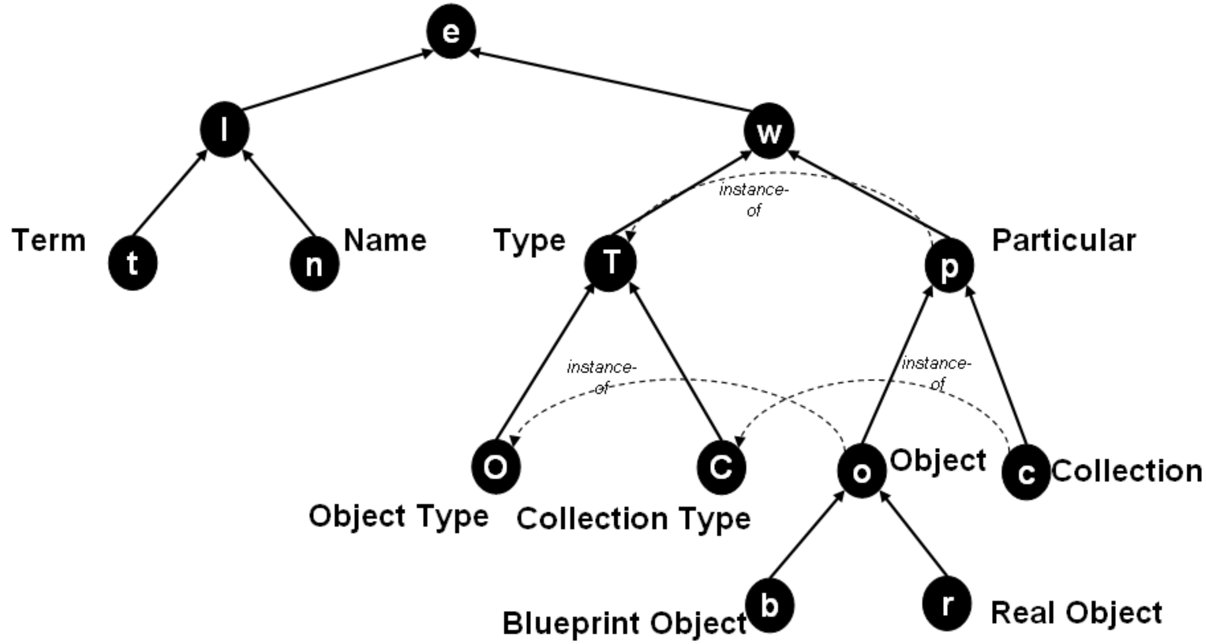




- These are different things!
  - "Human insulin has a molecular weight of 5,808 Dalton"
    Can be said about 2 only
  - "Human insulin is a crystalline white substance":
    Can be said about 3 only.
- General observation: molecules / genes are frequently seen as instances, i.e. reference to "blueprint" objects (or information entities)

Term / Name / Entity: Taxonomic Framework





- "[HDL]$^{n:O}$ has the ability to facilitate [transport of cholesterol]$^{t:O}$ to the [liver]$^{t:O}$."
- "[DACH1]$^{n:O}$ negatively regulates the [human RANK ligand gene]$^{n:O}$ expression in [stromal cells]$^{t:C}$."
- "[Carbon monoxide concentrations]$^{t:C}$ in [Santiago City]$^{n:r}$ at street levels..."
- "(...) [asthma]$^{t:O}$ severity reveal new regions of linkage in [EGEA study]$^{n:r}$ families."
- "Evaluation of two formulations of adjuvanted [RTS, S malaria vaccine]$^{n:C}$ in children (...) living in a [malaria]$^{t:O}$-endemic region of [Mozambique]$^{n:r}$."
- "In the current version of the [Mars Program Plan]$^{n:b}$, the [Astrobiology Field Laboratory]$^{n:b}$ ([AFL]$^{n:b}$) exists as a candidate project to determine whether there were (or are) habitable zones..."

Criteria for name / term distinction

- Ontological criteria are not sufficient for term / name distinction.
- Why is “liver” a term and “HDL” a name?
- Pragmatic criteria for term / name distinction:
  - Terms
    - Collected in dictionaries / domain terminologies
    - Stable, standardized
  - Names
    - Not systematized
    - Instable, dynamic, ad-hoc
  - Names may become terms

Tentative clarification

- Language level:
  - Name
    - Proper name
    - Term
- World level
  - Lexicalized entity
  - Non-lexicalized entity

Name / Proper Name

- A name is a unity of written human language denoting some entity (in the broadest sense). Names can be created and given by anybody familiar with this domain. The only name-forming rule is the use of accepted symbols.
- A proper name is a name which denotes a particular entity, i.e., an entity that cannot be instantiated (e.g., a person, a company, a brand, an institution or a geographic location)

Term

- A term is a name which is frequently used within a community.
- It has an unambiguous meaning within the user community.
- Terms are listed in terminology systems and dictionaries.
- Terms always denote types (?).
- Names which denote types can become terms provided the above mentioned criteria are fulfilled.

Lexicalized Entity

- Entity in the world (in the above sense) denoted by a unit of language (a word or a term) for which a documented shared meaning exists.
- Examples:
  - *T4 Lymphocyte* denotes a specific type of blood cell,
  - *Human Insulin* denotes either a type of peptide molecules or the particular blueprint of a peptide molecule.
  - *Dagstuhl* and *Boehringer Ingelheim* denote particular entities,

Nonlexicalized Entity

- Entity of the world referred to by units of language for which no commonly shared lexical entry exists
- Non-lexicalized entities are of high interest in emerging sciences
- Target of NER techniques are used to identify and relate new, non-lexicalized names.
- Smooth transition between lexicalized and non-lexicalized entities, in the sense that the latter get gradually accepted and re-used within a scientific community over time.

Conclusion

- Advanced NLP needs a consistent meta-terminology
- Cross-discipline communication
- Candidates for clarification:
  - (Named) Entity
  - Concept
  - Class
  - Term
  - Name
  - Ontology
  - Knowledge
  - …