An ontological modeling approach to neurovascular disease study: the NEUROWEB case

G. Colombo, D. Merico, **G. Frisoni**, M. Antoniotti, F. De Paoli, G. Mauri

Università degli Studi di Milano-Bicocca
Department of Information and Communication Technology (DISCo)

NEUROWEB Project
EU Sixth Framework Program
(Integrated biomedical information for better health)

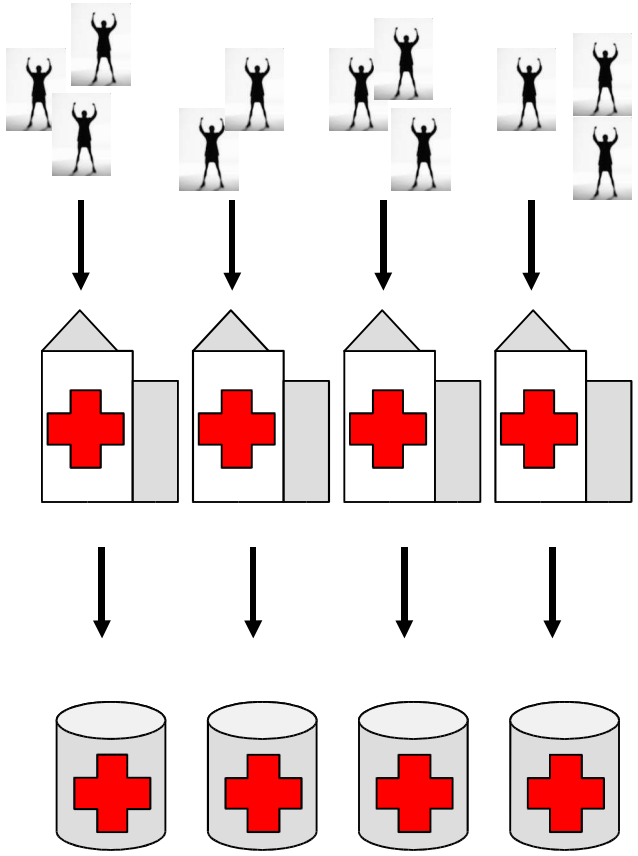
Presentation outline

- NEUROWEB Project
  - Project aims
  - Emerging issues
- The strategy adopted for ontological modeling:
  - Integration and ontological problems
  - The Knowledge Acquisition campaign
  - The Reference Ontology architecture
- The Reference Ontology structure
  - The Top Phenotypes: a stroke classification system
  - The Low Phenotypes: modular building blocks
  - An example of phenotype definition

The NEUROWEB Project: Aims

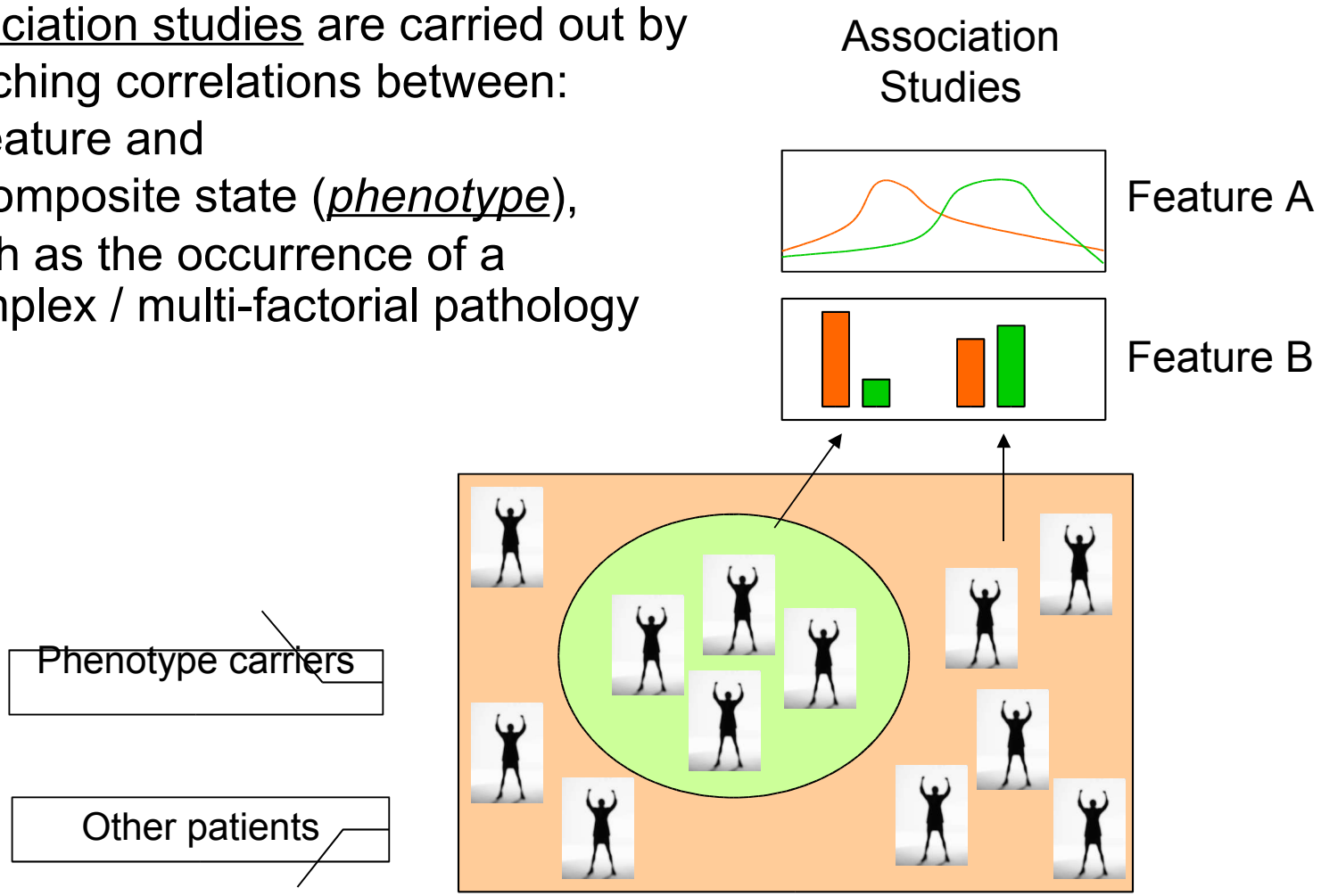
- NEUROWEB Aims:
  - support genomic association studies in the field of neurovascular medicine
  - provide a data integration framework for the participating clinical institutions

- NEUROWEB partners:
  - 4 EU clinical institutions being recognized excellence centers for stroke treatment
  - Each center makes available his clinical repository to other partners
  - The repositories store the results of clinical exams performed to reach a refined stroke diagnosis



The NEUROWEB Project: Aims

Association studies are carried out by searching correlations between:
- a feature and
- a composite state (*phenotype*), such as the occurrence of a complex / multi-factorial pathology



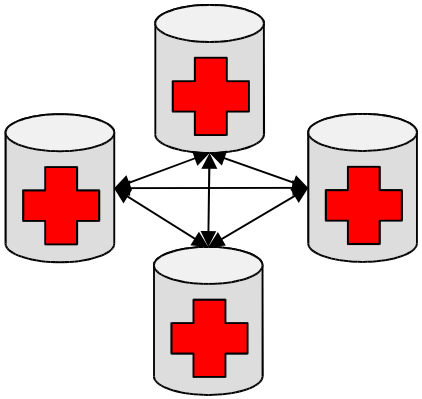
The NEUROWEB Project: Aims

- Association studies are carried out by searching correlations between:
  - a feature and
  - a composite state (*phenotype*), such as the occurrence of a complex/multi-factorial pathology
- Correlations can be imported from public genomic databanks
- In genomic databanks phenotypes are different (granularity, aim, etc.) from clinical phenotypes.
- The NEUROWEB Reference Ontology is conceived as the bridge between the clinical and the genomic phenotypes

The NEUROWEB Project: Issues

Association studies require the largest possible patient cohorts

Use data from different clinical sites



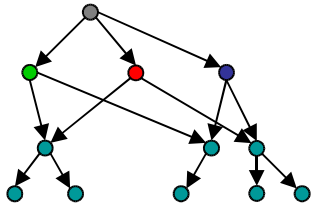
Data Integration problem

Clinical data collected during the diagnostic process are stored in repositories, designed according to local standards

deeply rooted in the expert knowledge of the local clinical community

Association studies require phenotype recognition

the occurrence of a *clinical phenotype* is asserted through the diagnostic process



Ontological problem: define phenotypes with a shared and explicit semantic

Data integration problem: heterogeneity

4 levels of heterogeneity in database integration:

- the *system level* → hardware and operating systems incompatibilities;
- the *syntactic level* → different DBMS;
- the *structural level*
  - data models
  - scales and measurement units
  - logic in grouping values (ranges)
- the *semantic level*
  - missing fields
  - one synthetic field vs. many analytical fields

Ontological problem: phenotypes with shared semantic

- In NEUROWEB the problem was not to find a common vocabulary to refer to shared meanings such as
  - use of the same term to mean different things;
  - use of different granularity to describe the same domain;
  - description of a domain from a different perspectives;
- …rather to find a shared meaning for well known terms (the phenotypes), such as "atherosclerotic ischemic stroke" or "lacunar stroke".
- We argued that each phenotype definition depends on
  - how the phenotype is observed
    - when, in respect of the stroke event
    - how the phenotype is measured
    - which device is used
    - where the phenotype is located in the body
  - the use of the phenotype
    - each local diagnostic and therapeutic process
- NEUROWEB needs a shared meaning for the phenotypes of interest based on the available data in each local database

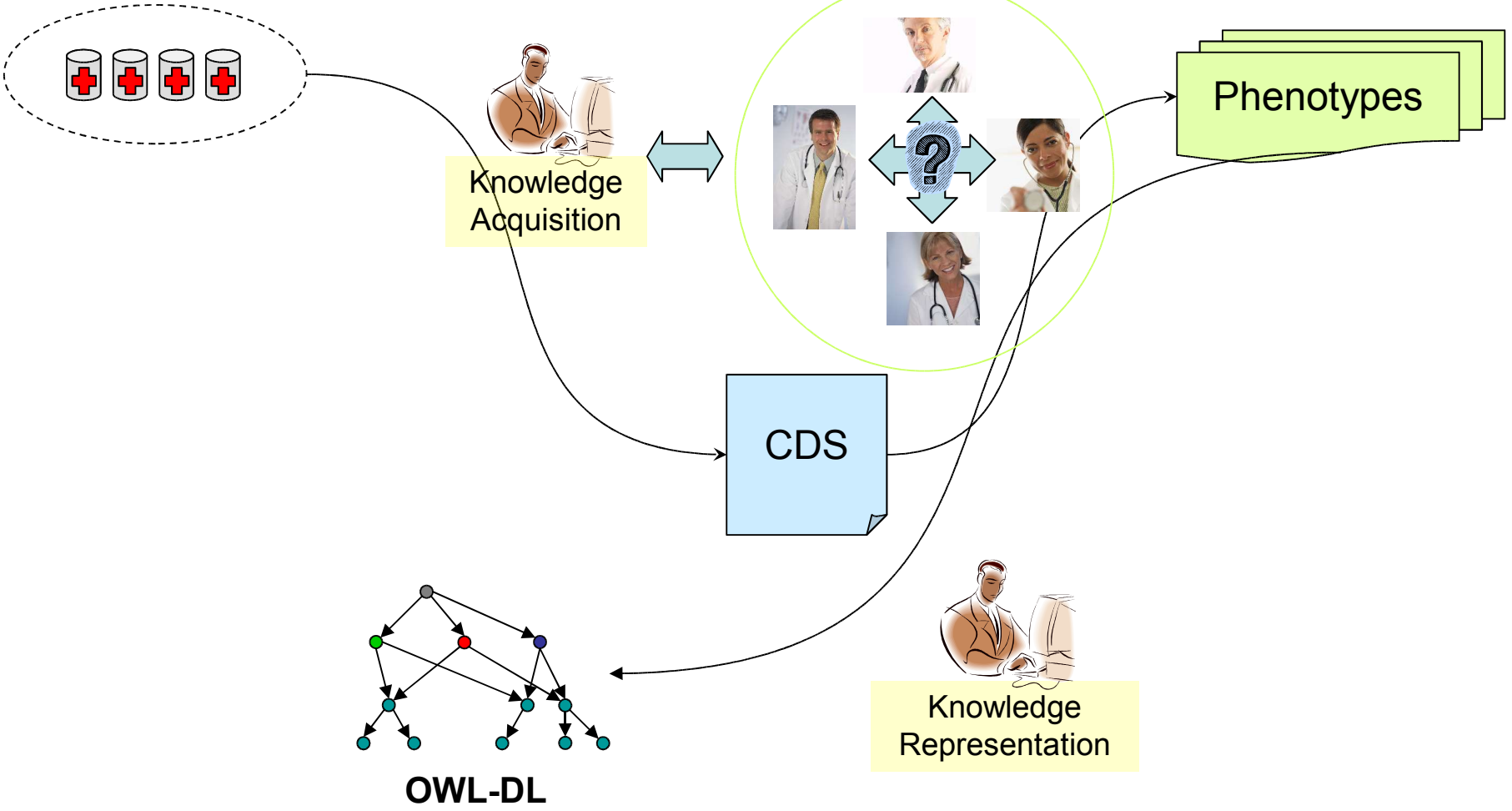
Ontological problem: phenotypes with shared semantic

"Categorization of subtypes of Ischemic Stroke has had considerable study, but definitions are hard to formulate and their application for diagnosis in an individual patient is often problematic."

Journal of the American heart association, *Classification of subtype of acute ischemic stroke*.

- We argued that each phenotype definition depends on
  - how the phenotype is observed
    - when, in respect of the stroke event
    - which device is used
    - how the phenotype is measured
    - where it is located
  - the local use of the phenotype
    - diagnostic and therapeutic process
- NEUROWEB needs a shared meaning for the phenotypes of interest based on the available data in each local databases

Ontology modeling strategy: the knowledge engineering approach



Ontology modeling strategy: the knowledge engineering approach

- Two major activities were carried out to produce the ontological model:
  - A major effort was done by clinicians to identify the straightforward similarities at the level of database content
    → generation of the **Core Data-Set** (CDS)
  - A Knowledge Acquisition campaign was carried out with the four medical centers, in order to identify the common set of phenotypes involved in the diagnostic process
    → generation of prototypal schemas for **phenotype** definition, exploiting the clinical profiles stored in each database

In turn, the analysis of these schemas revealed that phenotypes are aggregate entities, which can be decomposed into modular building blocks

The Reference Ontology

- Clinical databases are usually:
  - made by software houses with few contacts with expert clinicians → not focused enough;
  - made by clinicians themselves → not efficient and reliable.
- Knowledge Acquisition campaign useful even for the definition of a new focused and reliable database schema → it comes from the interaction between expert clinicians and technicians.
- The Reference Ontology is based on a set of data that clinicians use daily (Core Data Set): so far the Reference Ontology has been “forced” to be grounded to the real needs of expert clinicians.

Data integration and Reference Ontology

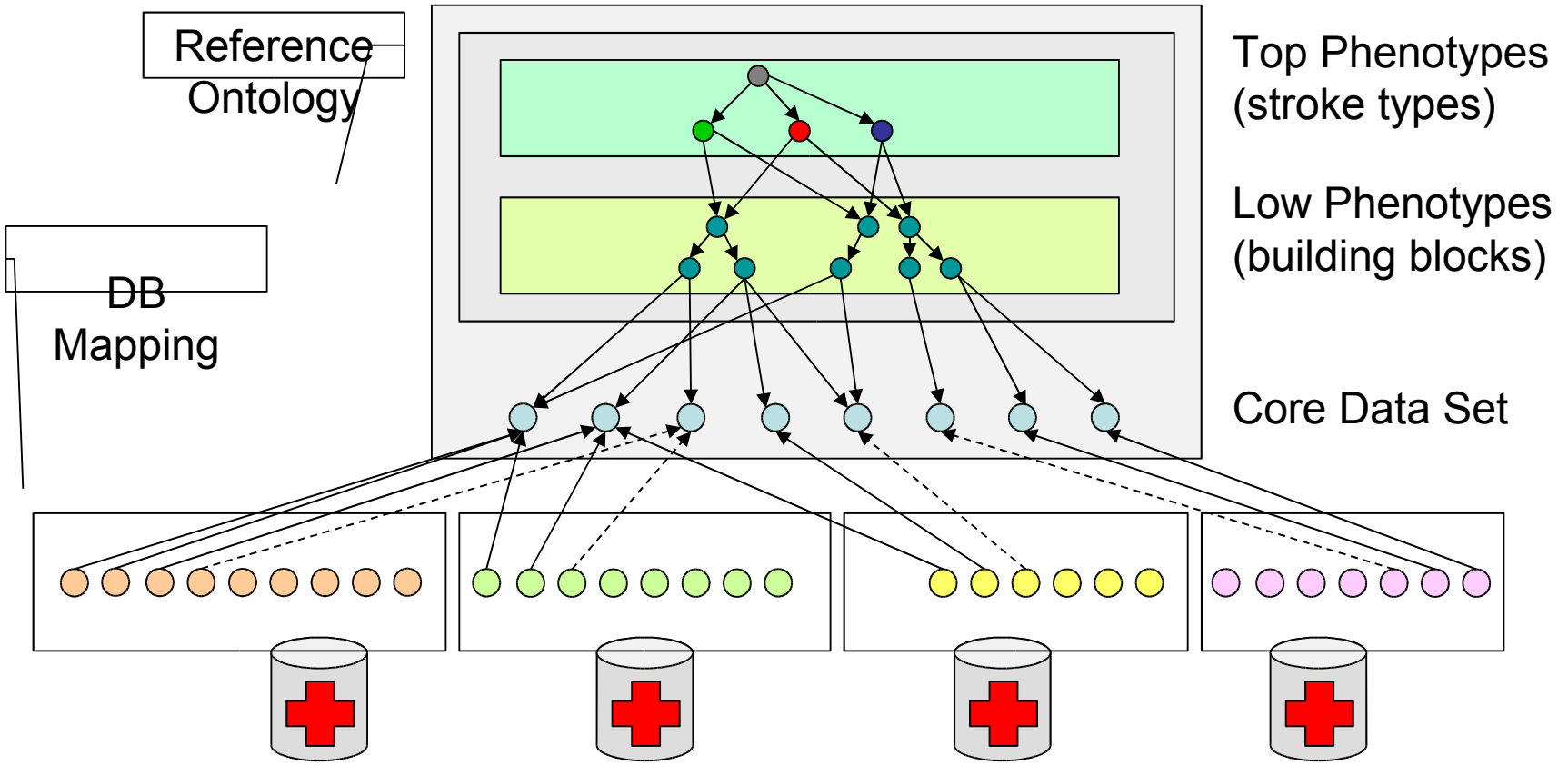
- The NEUROWEB Reference Ontology is both:
  - an issue to be faced in itself:
    - ontological problem in the knowledge engineering field and
  - a way to simplify the semantic level of the integration issue:
    - one synthetic field vs. many analytical fields → definition of a set of **shared synthetic fields**, called the Core Data Set (CDS).

The Reference Ontology
why a brand-new ontological model?

- What are the reasons why we did not adopt an already developed ontology?
  - phenotype ontologies in the genomic field are not suitable for *clinical* concepts
  - generalist medical ontologies are not committed to *phenotype* representation for association studies
  - generalist ontologies could prove unsuitable to represent the specificities of the expert knowledge characterizing the local neurovascular communities

Ontological modeling strategy ontology architecture

- The NEUROWEB Ontological framework manages both the data integration problem and the shared phenotype definition problem



The Reference Ontology
The Top Phenotypes Layer

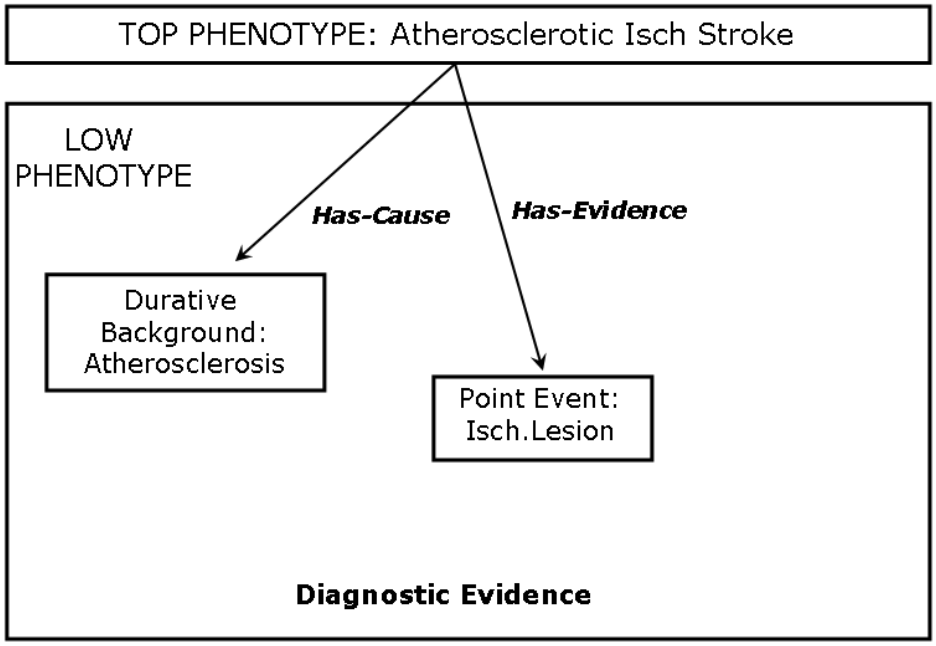
- The Top Phenotypes layer is a taxonomy of stroke types (e.g. *Atherosclerotic Stroke*) and related disease types (e.g. *Subclinical Atherosclerosis*), which is specifically adherent to the diagnostic procedures of the NEUROWEB clinical centers
- In this layer, phenotypes are seen just as labels allowing to classify a group of patients under it, in order to perform association studies; they are inter-related by IS-A relations
- The aggregate nature of phenotypes is taken into account by the underlying layer, the Low Phenotypes, which can be used to build new Top-Phenotypes in a modular process
- The connection between the Low Phenotypes and the Core Data-Set allows to root a Top Phenotype definition on the clinical repository content

The Reference Ontology
The Low Phenotypes Layer

- *Top Phenotypes* are decomposed into *Low Phenotypes*, through two main relations:
  - ***Has-Cause***,

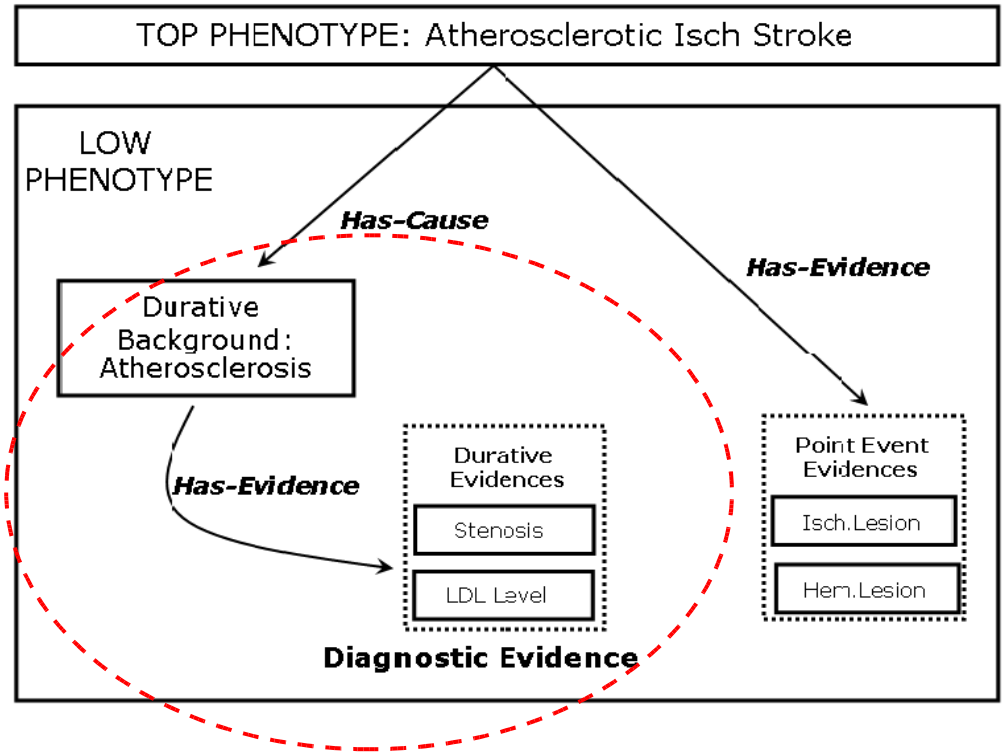
    pointing to the pathological process providing the *durative* etiological background for the stroke (i.e.: *Atherosclerosis*);
  - ***Has-Evidence***,

    pointing to the morphological evidences (i.e.: *Ischemic Lesion*) for the *point-events* leading to stroke.



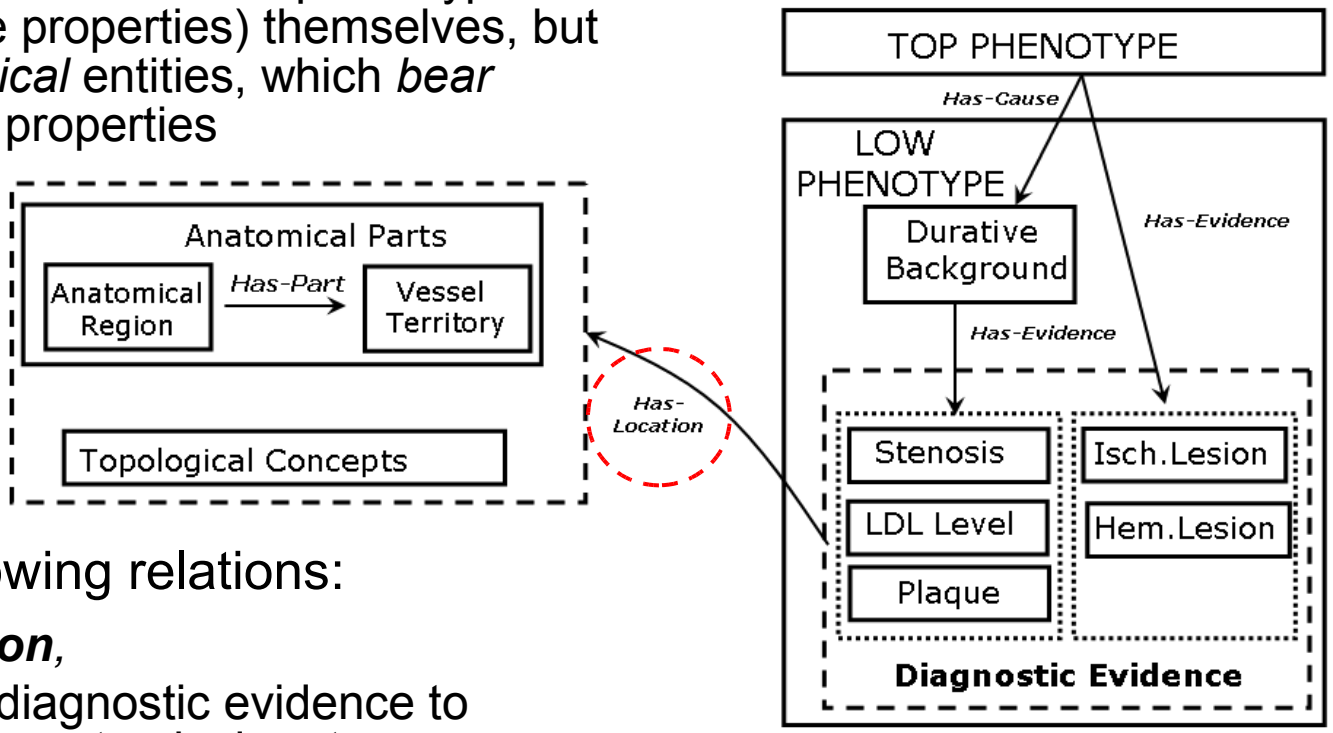
NEUROWEB Ontology: contents and structure overview

- The durative background is often a systemic disease (i.e.: atherosclerosis, diabetes), which cannot be directly observed, but instead requires an array of diagnostic evidences to be recognized; therefore, it is connected through the relation:
  - *Has-Evidence,* pointing to its diagnostic evidences (i.e.: *Stenosis, LDL Level*).



NEUROWEB Ontology: contents and structure overview

- Low Phenotypes are also connected to *Anatomical Parts*
  - *Anatomical Parts* are not phenotypes (observable properties) themselves, but rather *physical* entities, which *bear* observable properties

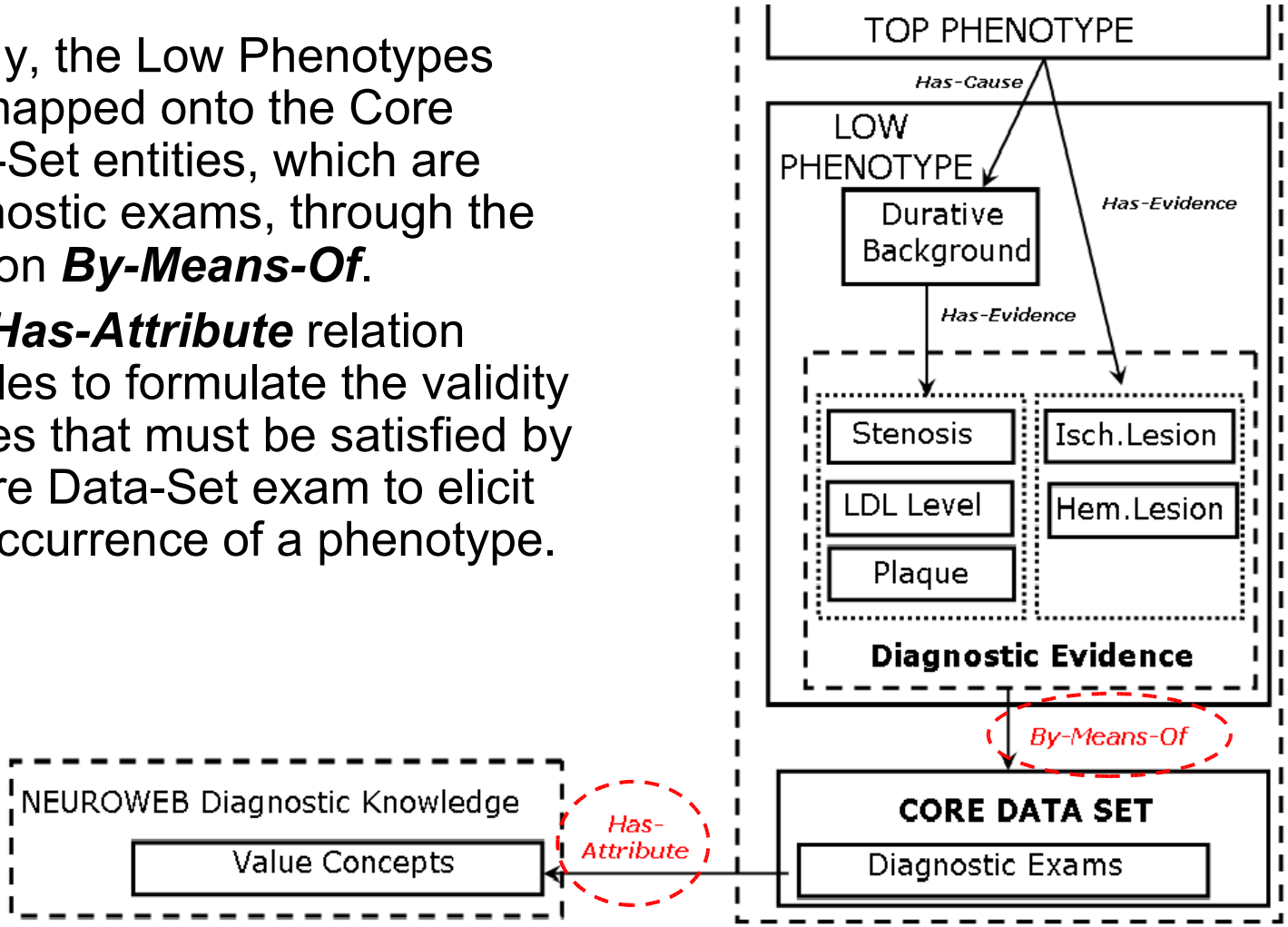


using the following relations:

- ***Has-Location***,
  connects a diagnostic evidence to the affected anatomical part;
- ***Has-Part***
  inter-connects anatomical parts.

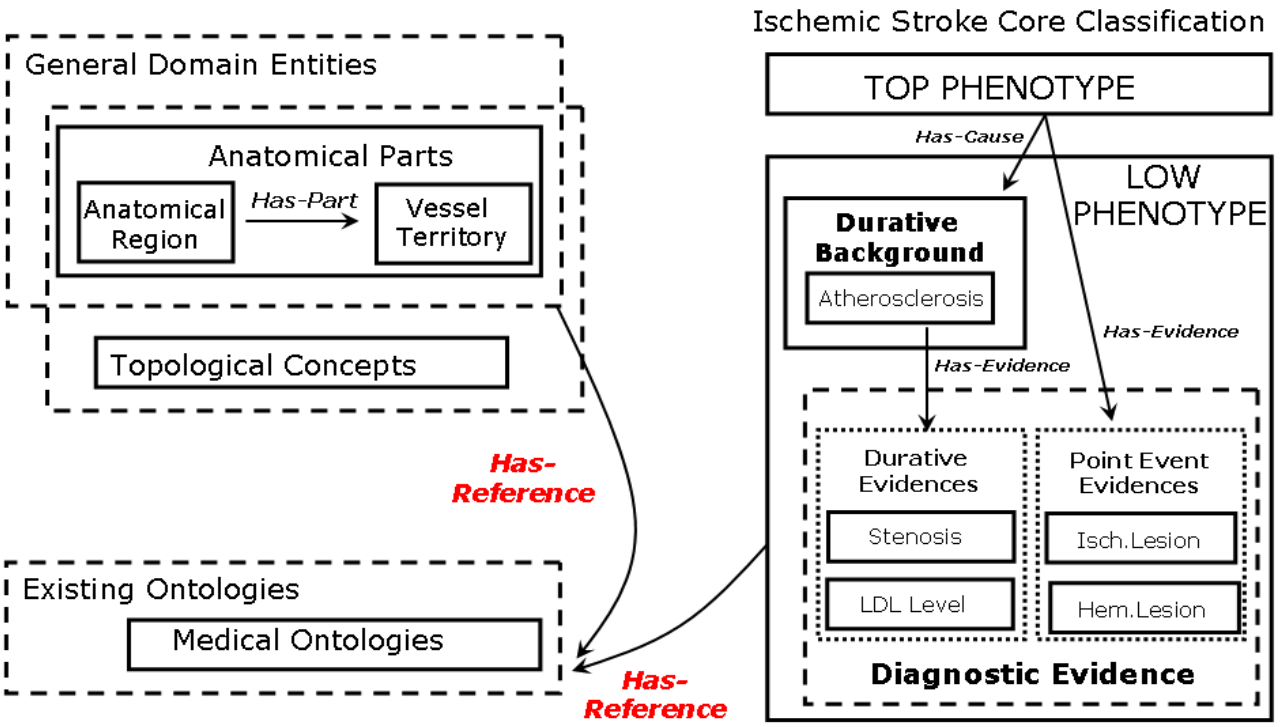
NEUROWEB Ontology: contents and structure overview

- Finally, the Low Phenotypes are mapped onto the Core Data-Set entities, which are diagnostic exams, through the relation ***By-Means-Of***.
- The ***Has-Attribute*** relation enables to formulate the validity ranges that must be satisfied by a Core Data-Set exam to elicit the occurrence of a phenotype.



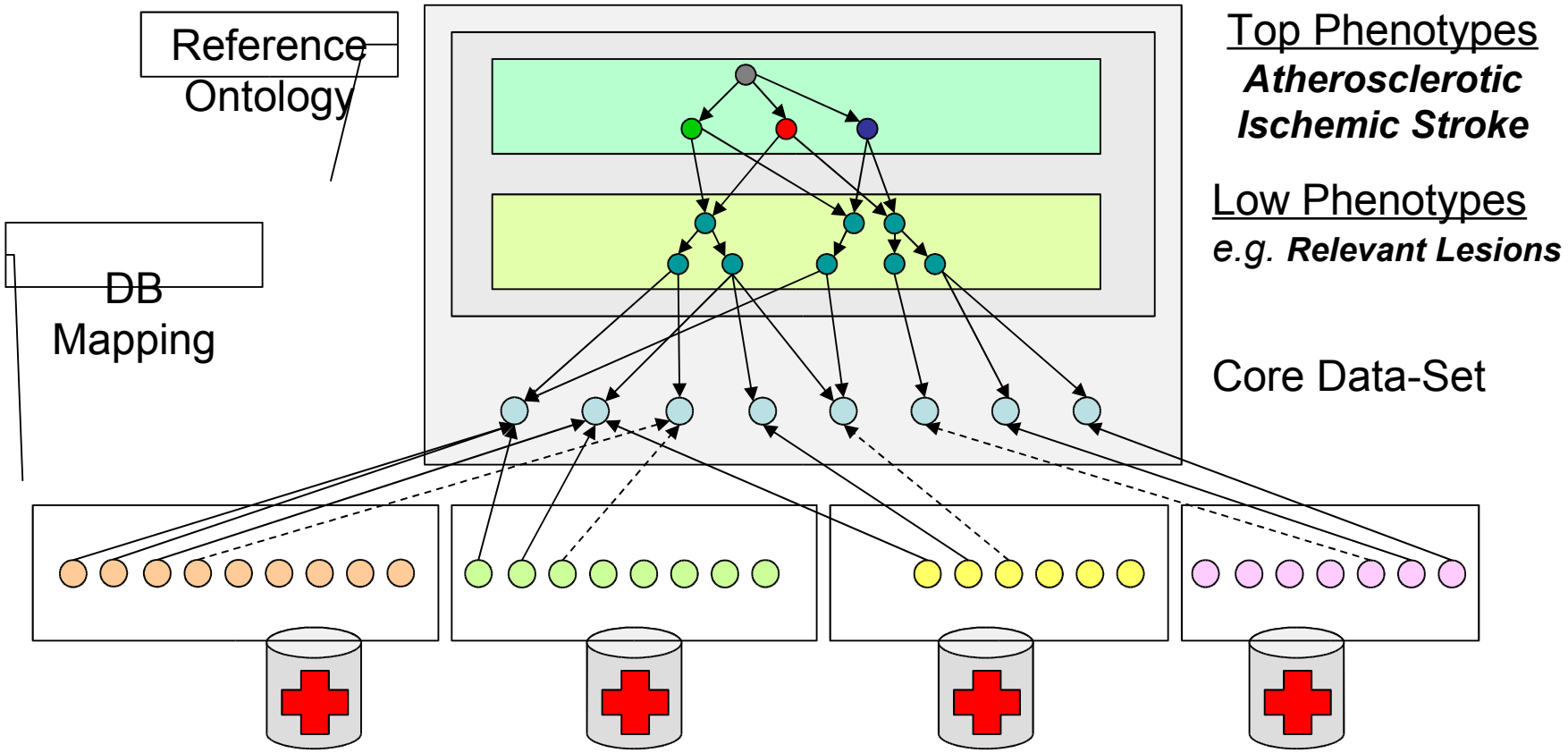
NEUROWEB Ontology: contents and structure overview

- The Reference Ontology is also mapped to other medical ontologies, in order to support queries on external resources:
  - At the present stage of development, we support integration with **SNOMED**, by linking Low Phenotypes and Anatomical Parts to corresponding SNOMED terms.

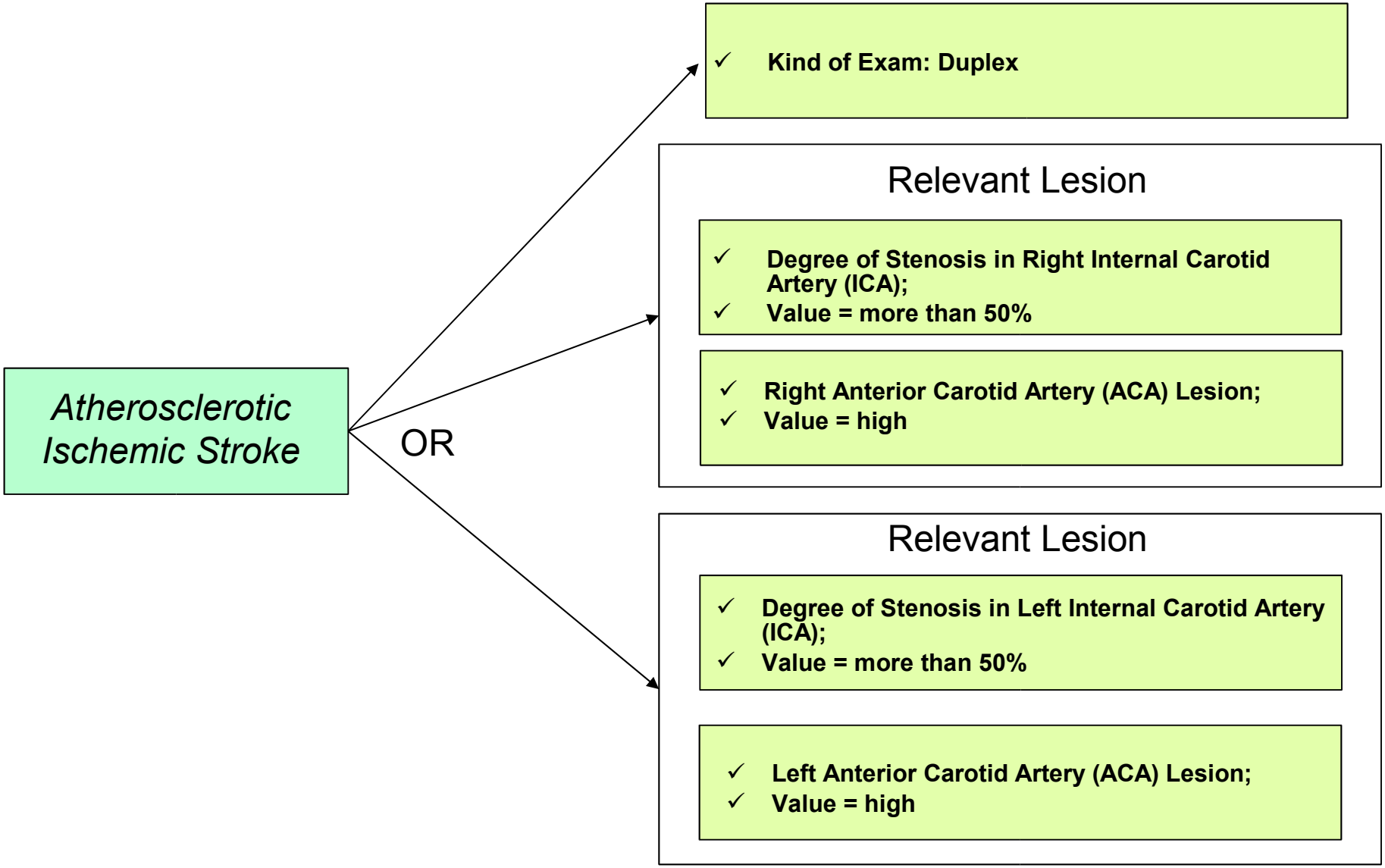


The Reference Ontology
An example of phenotype definition

- Phenotype: Atherosclerotic Ischemic Stroke;
- Clinical data to be used (a fragment of the required exams in order to validate it):



The Reference Ontology
An example of phenotype definition



The Reference Ontology
An example of phenotype definition

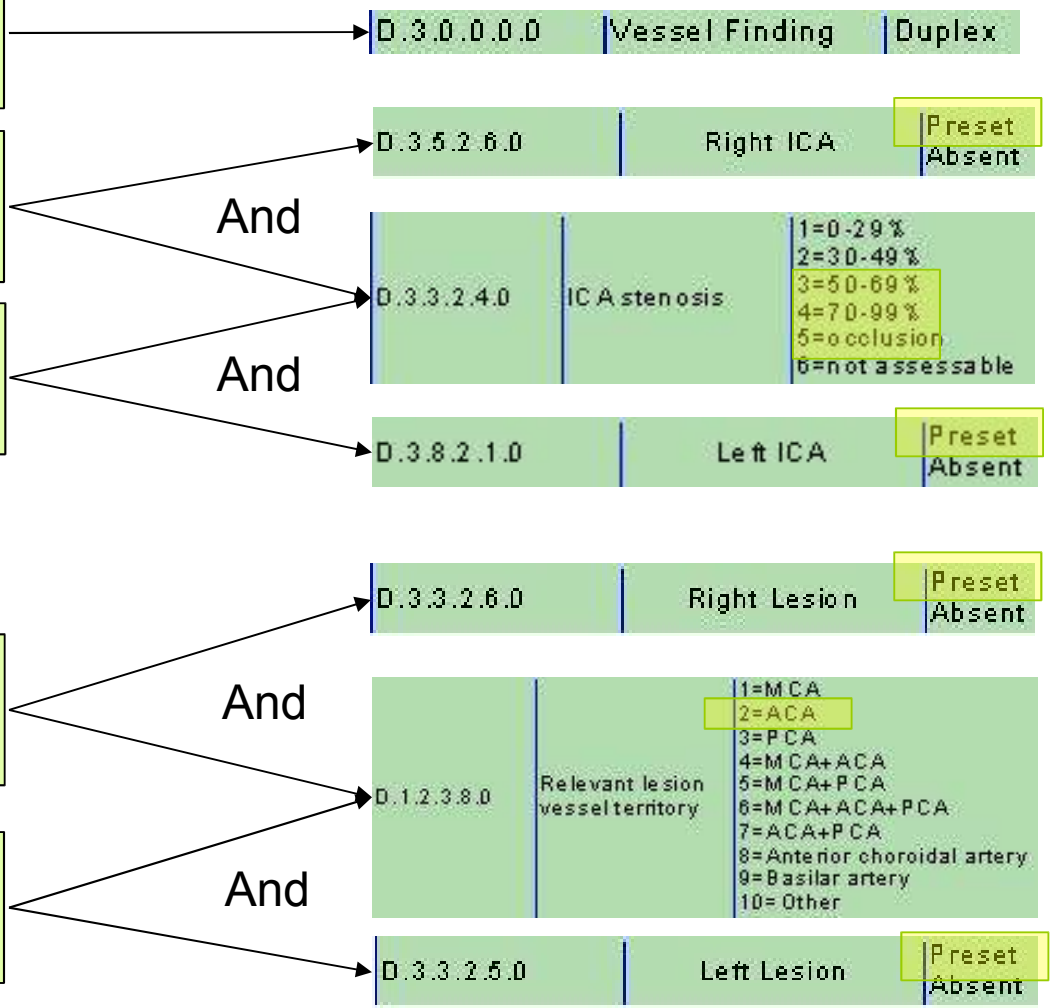
- ✓ **Kind of Exam: Duplex**

- ✓ **Degree of Stenosis in Right Internal Carotid Artery (ICA);**
- ✓ **Value = more than 50%**

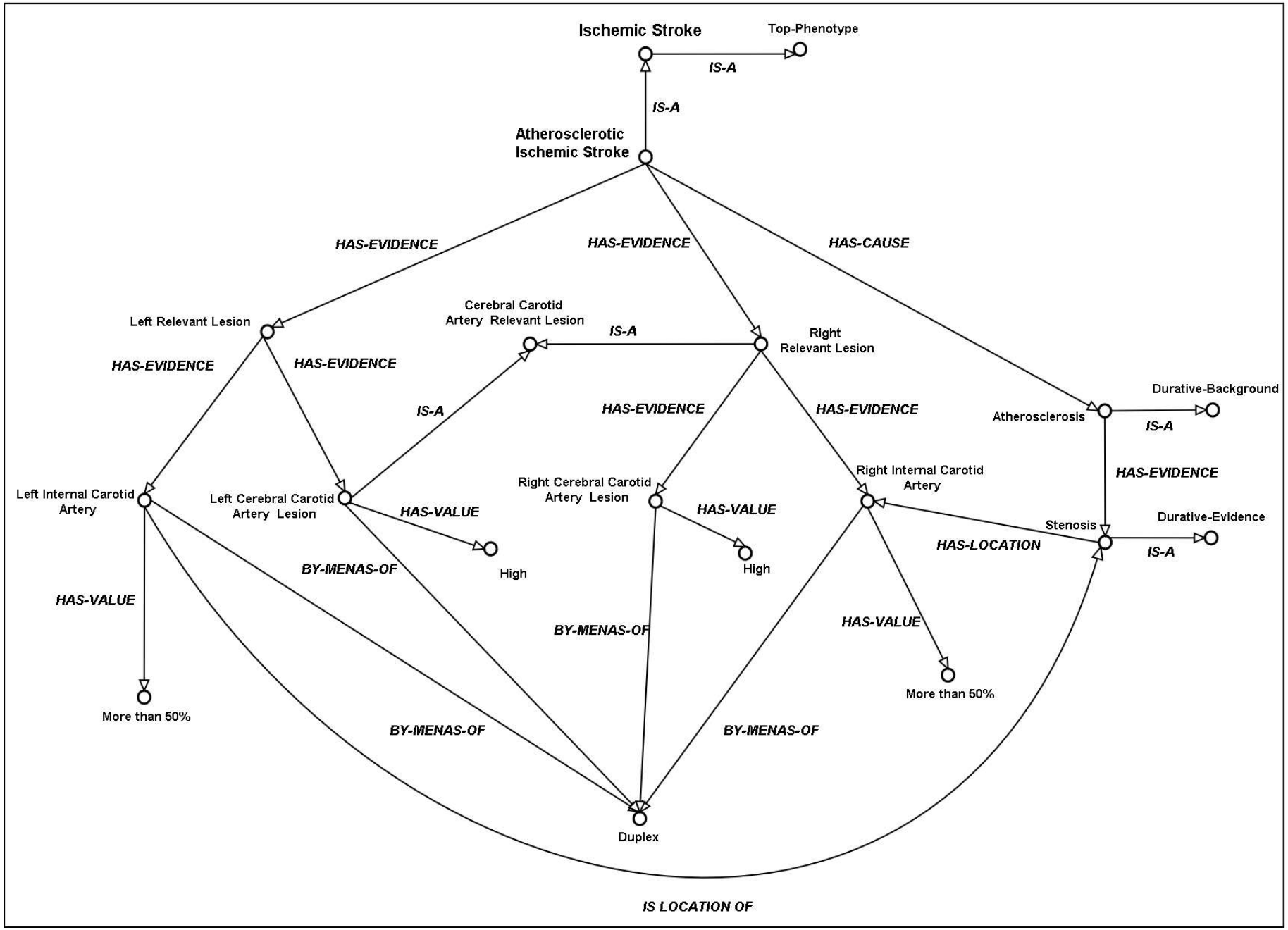
- ✓ **Degree of Stenosis in Left Internal Carotid Artery (ICA);**
- ✓ **Value = more than 50%**

- ✓ **Right Anterior Carotid Artery (ACA) Lesion;**
- ✓ **Value = high**

- ✓ **Left Anterior Carotid Artery (ACA) Lesion;**
- ✓ **Value = high**



NEUROWEB REFERENCE ONTOLOGY



The Reference Ontology
An example of phenotype definition

- In this way all the onto-logic formulas represent the instructions for the correct building of a complex phenotype as in the following:

$$\begin{aligned}&\mathtt{AtheroscleroticIschemicStrokeEvident} \sqsubseteq \\ &(\exists \mathtt{hasEvidence}.(\mathtt{LeftRelevantLesion} \sqcap (\mathtt{hasEvidence}.\mathtt{LeftICA} \sqcap (\exists \mathtt{isLocationOf}.\mathtt{Stenosis}) \sqcap \\ &(\exists \mathtt{hasValue}.\mathtt{Morethan50}) \sqcap (\exists \mathtt{byMeansof}.\mathtt{Duplex})) \sqcap \exists \mathtt{hasEvidence}.\mathtt{LeftCCALesion} \sqcap \\ &(\exists \mathtt{hasValue}.\mathtt{High}) \sqcap (\exists \mathtt{byMeansof}.\mathtt{Duplex})) \sqcup \mathtt{hasEvidence}.(\mathtt{RightRelevantLesion} \sqcap \\ &(\mathtt{hasEvidence}.\mathtt{RightICA} \sqcap (\exists \mathtt{isLocationOf}.\mathtt{Stenosis}) \sqcap (\exists \mathtt{hasValue}.\mathtt{Morethan50}) \sqcap \\ &(\exists \mathtt{byMeansof}.\mathtt{Duplex})) \sqcap \exists \mathtt{hasEvidence}.\mathtt{RightCCALesion} \sqcap (\exists \mathtt{hasValue}.\mathtt{High}) \sqcap (\exists \mathtt{byMeansof}.\mathtt{Duplex}))) \sqcap \\ &\exists \mathtt{hasCause}.(\mathtt{Atherosclerosis} \sqcap (\exists \mathtt{hasEvidence}.\mathtt{Stenosis}))\end{aligned}$$

Conclusions and Future Works

- We have developed an ontological framework providing:
  - A robust but flexible representation of clinical phenotypes, in order to support phenotype-genotype association studies;
  - A phenotype definition rooted onto the diagnostic process, which mirrors the mental scheme by which clinicians analyze and understand disorders;
  - A phenotypes representation as aggregates of building-blocks, so that the already defined ones can be customized by removing or adding discrete components.
- The Ontology has been implemented via OWL-DL.
- We are working on the system computational architecture in order to exploit Ontology for:
  - The DBs integration;
  - As the enabling factor of the NEUROWEB functionalities;
  - The user interface definition.

References

- Bard J.B.L., Rhee S.Y. (2004) Ontologies in biology: Design, applications and future challenges. *Nature Reviews Genetics,* 6(5), 213-222.
- Bodenreider, O., Stevens, R. (2006) Bio-ontologies: current trends and future directions. *Brief. Bioinform.*, 7(3), 256-274
- Rector, A.L., Rogers, J.E. (2005) *Ontological and Practical Issues in Using a Description Logic to Represent Medical Concepts*. Computer Science PrePrint, University of Manchester.
- Consortium TIH. (2005) A haplotype map of the human genome. *Nature,* 437 (7063), 1299-1320
- Botstein D., Risch N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, 33, S228-237.
- The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., 25, 25-29.
- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, 34, D354-357.
- Reactome [http://www.reactome.org]
- Ay H., Furie K.L., Singhal A., Smith W.S., Sorensen A.G., Koroshetz W.J. (2005) An evidence-Based Causative Classification System for Acute Ischemic Stroke. *Ann. Neurol.,* 58, 688–697.
- W3C [http://www.w3.org/TR/owl-guide/].
- Sattler U, Description Logics for the Representation of Aggregated Objects. In *Proceedings of the 14th European Conference on Artificial Intelligence*. Edited by W. Horn, IOS Press, Amsterdam 2000.
- Adriana S. Apar, Oscar L. M. Farias and Neide dos Santos, Applying ontologies in the integration of heterogeneous relational databases. *In Proceedings of the 2005 Australasian Ontology Workshop*, 2005
- Adams HP, Love BB, Gordon DL, Marsh EE, Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Journal of the American Heart Association, 24, 35-41, 1993

The Reference Ontology vs SNOMED