Integrating an Italian consumer health terminology with the UMLS using Semantic Web Technologies

Elena Cardillo
cardilloe@mail.nih.gov

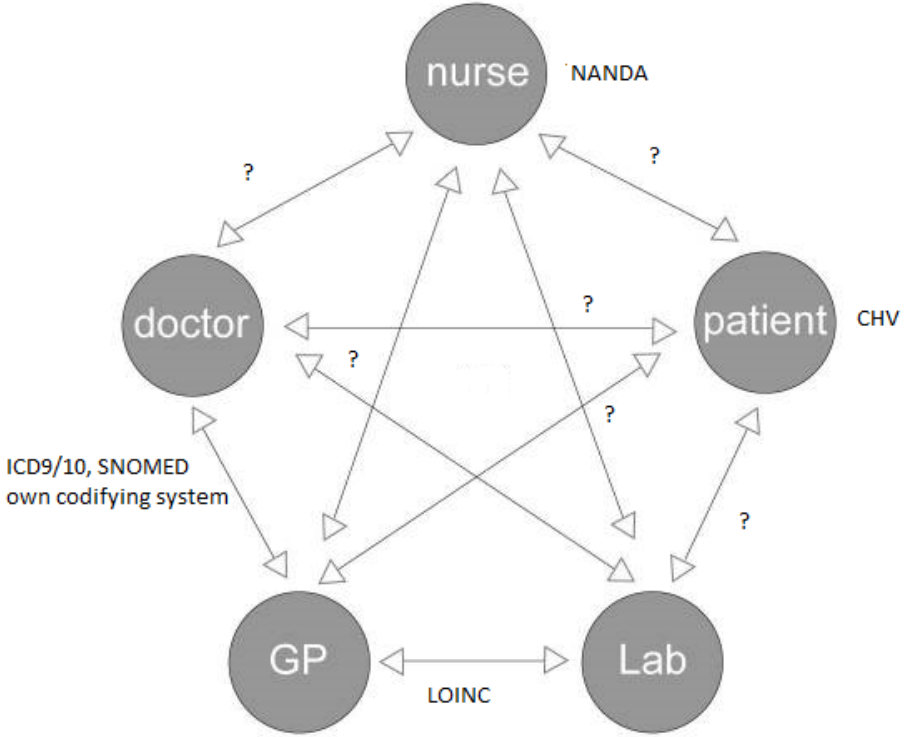
Supervised by: Olivier Bodenreider

Introduction

- Huge effort in integrating medical terminologies and classification systems by creating mappings between them
- Use of Semantic Web Technologies
  - To formalize existing medical terminologies
  - To develop new medical ontologies
  - To integrate them into large ontology repositories (e.g. BioPortal)
- More emphasis on the patient perspective
  - Personal Health Records accessible from the web
  - Active role played by consumers
  - Development of consumer-oriented vocabularies

Critical Issues

- Interoperability between different healthcare systems is still a significant problem
- Medical ling
  - Differenc professionals
    - Epistaxi
    - Dyspnea
- Need for int professional medical voc ormation systems suc



Objectives

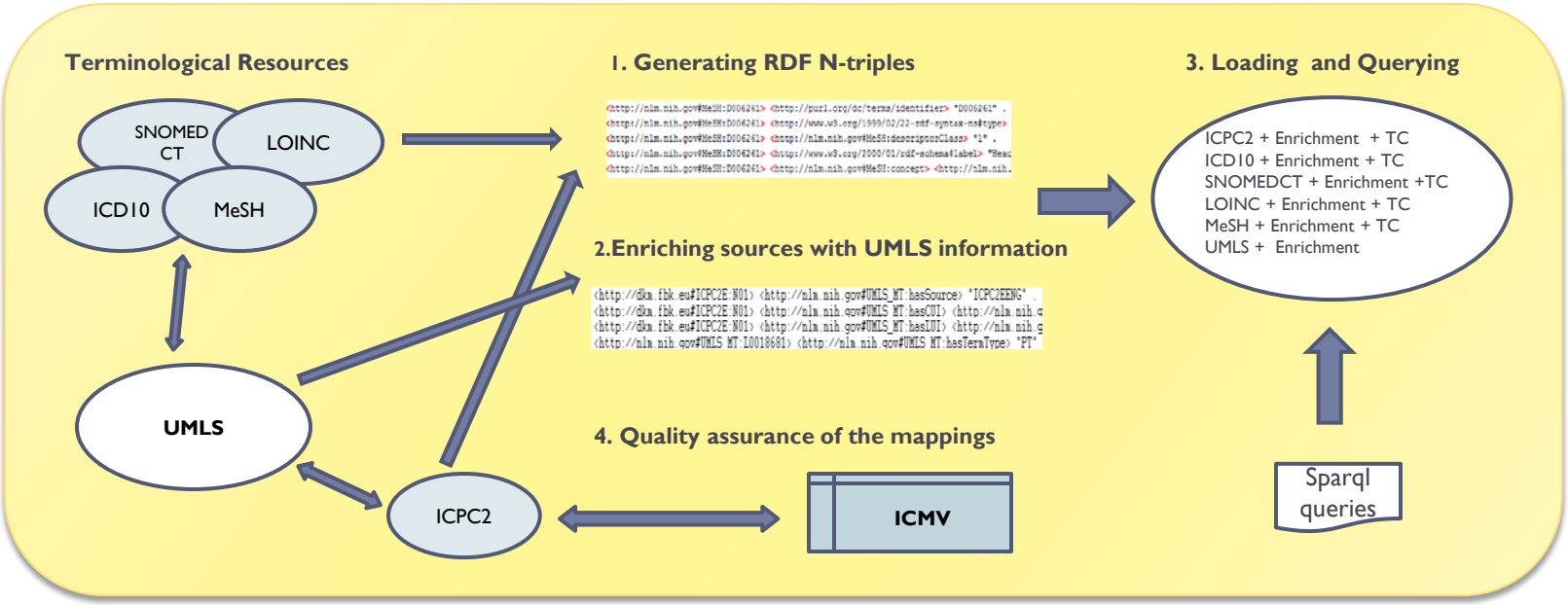
- To create an Integration Framework for the General Practice domain
  - Map language-dependent consumer-oriented vocabularies to language-independent professional medical terminologies
- Why:
  - To help bridge the linguistic gap between lay and professional resources
  - To help consumers interpret clinical notes and test results and describe clinical history and complaints in their PHRs
  - To facilitate querying and searching of healthcare information
  - To improve consumer-oriented healthcare information systems
- How:
  - Using UMLS as a source of mappings between medical terminologies
  - Using Semantic Web technologies for integration purposes

Materials

- Consumer health terminology
  - Italian Consumer Medical Vocabulary (ICMV) – v. 2009
    - 1659 Italian lay terms for symptoms, diseases and anatomical structures, 1355 mapped to ICPC2
- Professional terminologies
  - SNOMED CT – RDF v. July 2009 (NLM ongoing project)
  - MeSH – RDF v. 2009 (NLM ongoing project)
  - LOINC – v. 2.27 2009
  - ICD10 – OWL v. 2008 (FBK ongoing project)
  - ICPC2 – OWL v. 2008 (FBK ongoing project)
- UMLS Methathesaurus - v. 2009AB
  - UMLS partial RDF version (NLM ongoing project)

Overview of the Methods

- ICPC2 serves as a pivot between ICMV and other professional vocabularies integrated in the UMLS
- UMLS Metathesaurus provides mappings between ICPC2 and SNOMED CT, MeSH, LOINC and ICD10



Approach:
Step 1. Generating RDF N-triples

- Medical terms and their inter-relations are represented using RDF N-Triples
  - <subject> <predicate> <object> .
- SNOMED CT and MeSH already converted to RDF for other NLM projects
  - <subject> a SNOMED CT concept or a MeSH descriptor
  - <predicate> concept properties
  - <object> a literal corresponding to a property or a node representing another concept
- OWL resources, ICPC2 and ICD10 serialized in RDF are directly compatible with other RDF resources
  - <subject> a class of the ontology (ICPC2 or ICD10 concept)
    - blank nodes for the representation of restrictions of the classes
  - <predicate> concept properties
  - <object> a literal corresponding to a property or a node representing another concept
- Java program to create RDF triples for LOINC from data in the UMLS Metathesaurus
  - Extraction of labels, type, and identifier for each LOINC "concept" and "part" from MRCONSO table
  - Extraction of relations among entities from MRREL table

Approach:
Step 2. Enriching sources with UMLS information

CUI I: C0009264 Cold Temperature

LUI I: L0215040	Cold Temperature	CSP
LUI 2: L0009264	Cold	MSH, MTH

CUI 11: C0009443 Common Cold

LUI I: L0009264	Cold	MTH, COSTAR
LUI II: L0009443	Common Cold	MSH

CUI III: C0024171 Chronic Obstructive Airway Disease

LUI I: L0498186	Chronic Obstructive Airway Disease	MSH
LUI II: L0009264	COLD	MSH, SNMI

…

Approach:
Step 3. Loading and Querying

- Loading N-triples into the Virtuoso RDF triple store
  - 17 graphs (6 Original resources, 6 UMLS Enrichment resources, and 5 Transitive Closure resources)
- 3 types of queries:
  1. Find concepts corresponding to ICPC2 concepts, using CUIs
  2. Find synonyms/new names corresponding to ICPC2 concepts, using CUIs, LUIs and TTYs
  3. Find common Parents relations among the terminologies, using CUIs
- SPARQL
  - Query language for RDF resources
  - Equivalent to SQL for relational databases
- Java program to automate the submission of batch queries to Virtuoso and collect the results

Approach:
Step 4. Quality assurance of the mappings

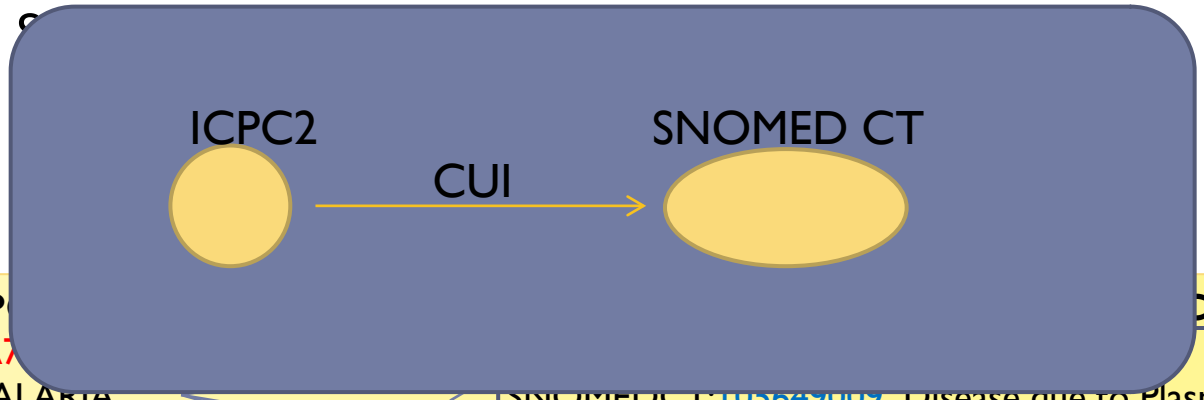
- Evaluation of the quality of the mappings between ICMV and ICPC2 and the suitability of ICPC2 for representing ICMV
  - Direct mapping of some "lay" ICMV terms to the Italian concepts in UMLS Methathesaurus
    - Using exact match (UMLSKS application programming interface)
      - Mapping found (ICMV term → UMLS CUI + Preferred Term + Source + Code)
      - No mapping found
  - Compare direct mappings through UMLS to the mapping through ICPC2 created by experts

Results: Triples and Performance

- Numbers of Triples loaded into Virtuoso
  - 66,769,781 unique triples among the 17 graphs, in particular:
    - 97,457 from ICD10
    - 18,650 from ICPC2
    - 1.9M from LOINC
    - 1.8M from SNOMED CT
    - 16.6M from MeSH
    - ~ 50M from UMLS
- Performance
  - Few seconds for loading each graph into Virtuoso, 5 minutes only for the UMLS Enrichment graph (larger UMLS graph already loaded)
  - Short execution time for batch queries (3-5 minutes for a query on each concept in ICPC2)
  - Poor performance for queries related to Hierarchical mappings (hours)

Results: Find concepts corresponding to ICPC2 concepts (CUI-based)

- S



N.of source Concepts

ICP

A7

MALARIA

Conc.

SNOMEDCT:105649009 Disease due to Plasmodiidae (disorder) [Parent]
SNOMEDCT:186797008 Unspecified malaria (disorder) [Inactive Concept]
SNOMEDCT:248437004 Malarial fever (finding) [Intermitted fever]

UMLS CUI: C0024530

ICD10 Conc.

F93

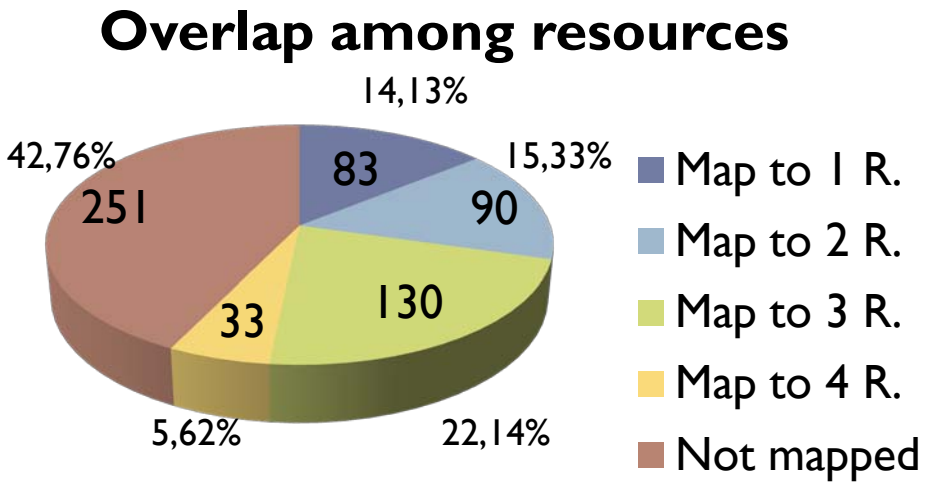
GLAUCOMA

ICD10:H40 Glaucoma
ICD10: H.40.9 Glaucoma, unspecified [Child]
ICD10:H40-H42.9 Glaucoma [Parent class]

UMLS CUI: C0017601

- Pairs of SNOMED CT concepts are collapsed in the same UMLS CUI. E.g. Malaria

Results: Find concepts corresponding to the ICPC2 concepts (CUI-based)

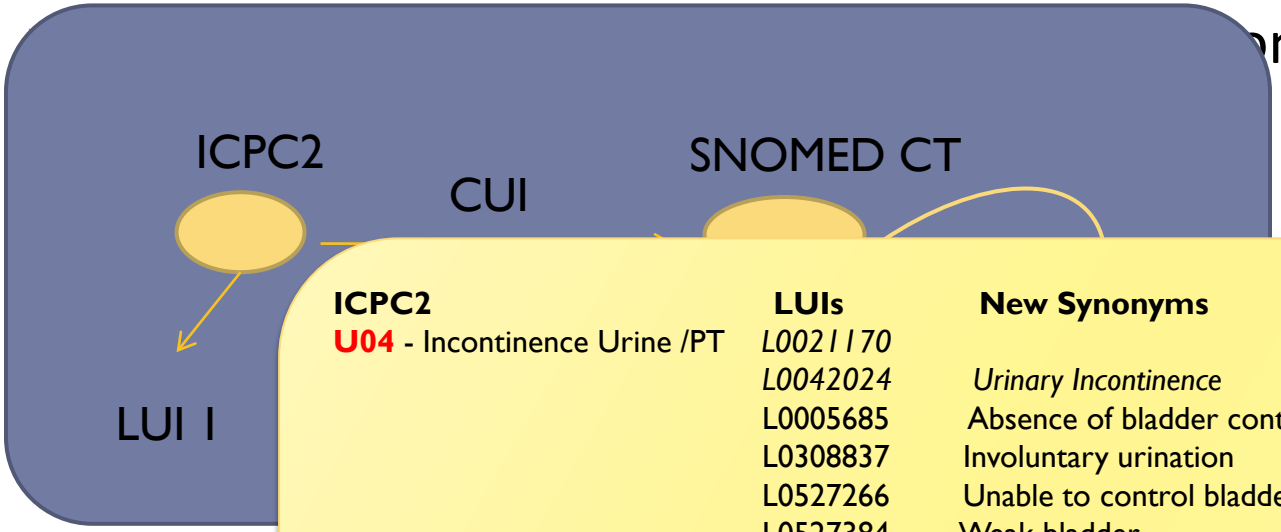


Examples
- Map to 4 terminologies:
  A03 - Fever, F93 - Glaucoma
- Map only to SNOMEDCT:
  A18 - Concern about appearance
- Map only to MeSH:
  N19 - Speech disorder
- Map only to ICD10:
  H77 – Sprain/strain of ankle

- Among the 83 mapped to only one terminology:
  - 74 map only SNOMED
  - 4 map only to MeSH
  - 5 map only to ICD10

Results: Find synonyms/new names corresponding to ICPC2 concepts

onding to 587



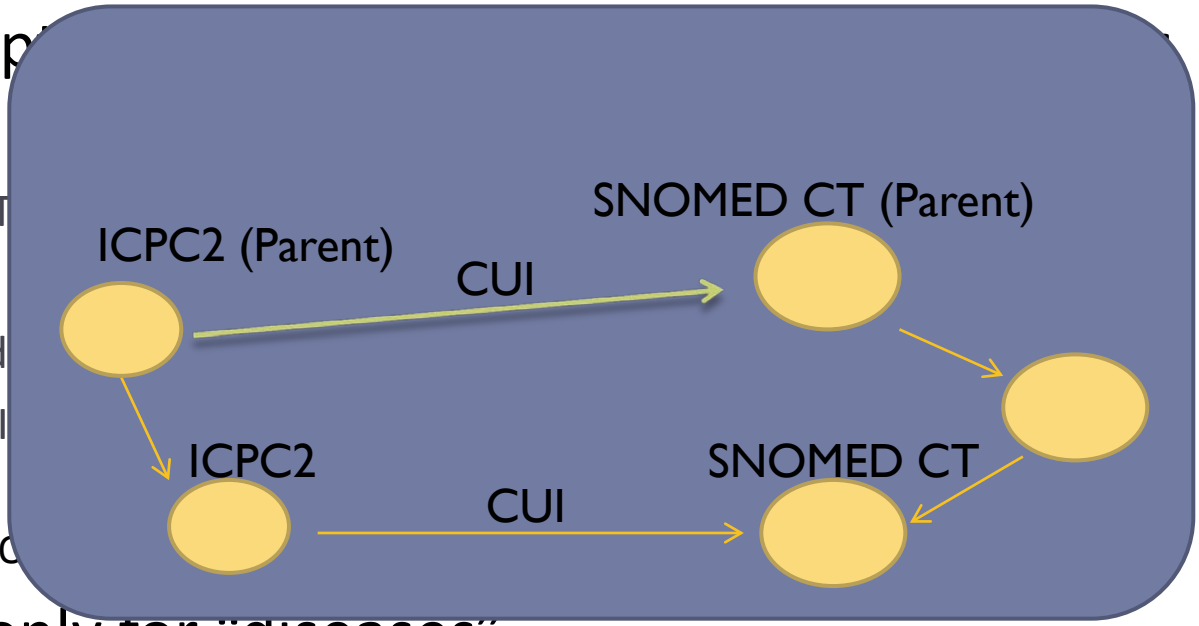
- 68 (62

- Synony
  - A tota
    - E.g.

ICPC2	LUIs	New Synonyms	SNOMED CT
U04 - Incontinence Urine /PT	*L0021170*		
	L0042024	*Urinary Incontinence*	
	L0005685	Absence of bladder continence	
	L0308837	Involuntary urination	
	L0527266	Unable to control bladder	
	L0527384	Weak bladder	
	L0583725	Unable to hold urine	
	L0005692	Bladder incontinence	
	L0527301	Unable to prevent bladder emptying	**165232002**
	L0574730	Unable to hold fluids	
	L0586619	UI - Urinary incontinence	
	L0748747	Bladder: incontinent	
	L0042024	Incontinence, urinary	
	L0527264	Lack of bladder control	
	L0527265	Loss of bladder control	
	L0590897	Leaking of urine	

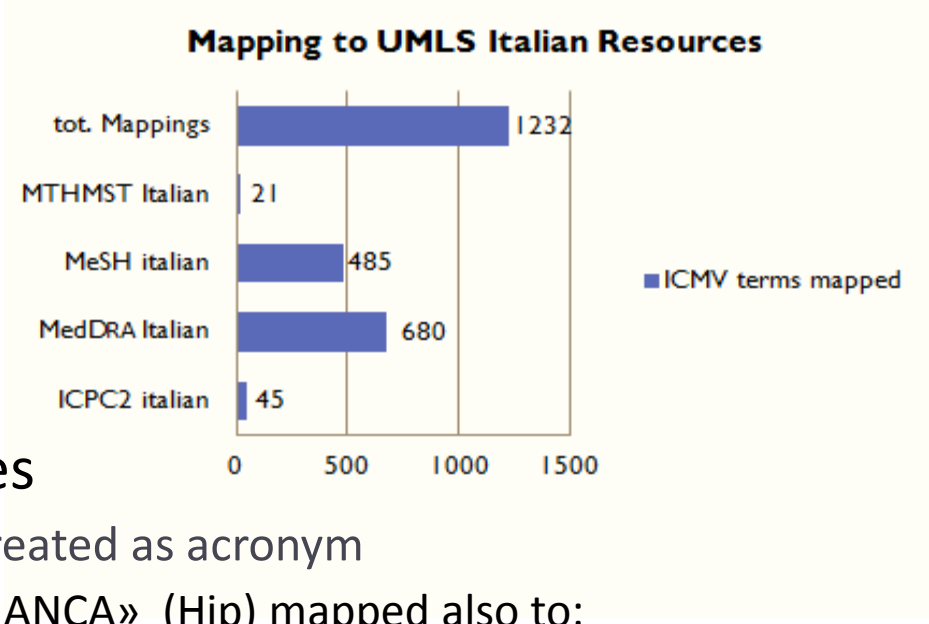
Results: Find common Parent relations among terminologies

- 220 ICPC2 concep
  terminologies
  - 193 with SNOMED CT
  - 27 with MeSH
  - None with ICD10 and
  - 84 unique parent CUI
    terminologies
    - 4/84 common to SNO
- Mapping found only for "diseases"
- Mappings with shared parents
  - 27/201 14% of ICPC2 concepts mapped to MeSH
  - 193/321 60% of ICPC2 concepts mapped to SNOMED CT



Results: Quality assurance of the mappings

- Mapping ICMV terms to UMLS Italian concepts (exact match)
  - 655/1659 unique ICMV terms
  - 1232/1659 total mappings to the UMLS Italian concepts
  - Mapped to 690 unique UMLS CUIs



- Ambiguity issues
  - Concept name treated as acronym
    - E.g. the term «ANCA» (Hip) mapped also to:
      - Anticorpi antineutrofili anticitoplasma (Antineutrophil Cytoplasmic Antibodies)

Conclusions

- ICPC2 integrated with SNOMEDCT, ICD10, MeSH, LOINC using RDF and SPARQL queries
  - 50% of ICPC2 concepts mapped to at least one other terminology
  - Many multiple mappings, that is "ambiguity"
  - Inconsistencies found in terms of classification of symptoms and diseases among the terminologies
    - E.g. "Warts" classified as symptom in ICPC2 and as disease or diagnosis in the other terminologies
- New mappings between ICMV and professional terminologies through our integration framework
- Future work
  - Classification of the ICMV terms according to the most representative professional terminology
  - Evaluation of the feasibility of this framework under the PHR at development at the FBK Research Institute in Trento (Italy)

Project outcomes

- One paper submitted to eHealth2010
  - Cardillo, E., Hernandez, G., Bodenreider, O.: Integrating consumer-oriented vocabularies with selected professional ones from the UMLS using Semantic Web Technologies.
- One in preparation for SMBM2010

Acknowledgment

- Thank you very much to NLM and LHC director and staff for this visiting period and in particular to:
  - *Olivier Bodenreider*
  - *May Cheh*
  - *Celina Wood*
  - *Genaro Hernandez*
  - *John Nguyen*
  - *Mohammed Cyclegar*
  - *All my colleagues and office neighbors*

Thanks for you for attention