Medical Concept Representation - the Years Beyond 2000

Laszlo Balkanyi[a], Stefan Schulz[b,c], Ronald Cornet[d,e], Olivier Bodenreider[f]

[a] European Centre for Disease Prevention and Control, Stockholm, Sweden, [b]Medical University of Graz, Austria, [c]Freiburg University Medical Center, Freiburg, Germany, [d]Academic Medical Center, Amsterdam, The Netherlands, [e]Linköping University, Linköping, Sweden, [f]National Library of Medicine, Bethsheda, USA

Abstract

- **understanding the state of the art** in the context of "**medical concept representation**"
- **a descriptive study** based on bibliometrics, simple text mining and a social media survey
- results support the general under-standing that the focus of research has moved toward medical ontologies
- socially active researchers mention the OBO foundry, SNOMED, and UMLS as key resources
- text mining of most cited literature identifies single noun phrases as "health", "information", "clinical", "knowledge", "ontology", "case", "data", " semantic(s)", "concept" and "representation" as leading denominators of the field
- terms as "**ontology**" and "**semantic(s**)" have gained more significance in the last decade
- **there is a paradigm shift** according to both the socially active group of researchers and bibliometric data, comparing citation ranks of the nineties and the recent decade support this opinion

Introduction

The goal of the exploratory study:

to understand the state of the art in the broad contextual research area of "**medical concept representation**" originating in the 1990s

Influencing factors:

- advances in medical information science,
- terminologies, ontology development,
- accessibility of networked computing

Effect:

- significant growth of research, development practical implementation in this area [1, 2].

Background:

The study was initiated by the IMIA Medical Concept Representation Working Group (MCR WG).

The WG was one of the most influential bodies in the late eighties and the nineties, publishing regularly overviews of the domain, last in-depth, analytic study dated to 2006 [3].

Current study based on:

- bibliographic measures,
- simple on-line text mining tools,
- a social media survey

Materials and Methods

Selecting sources of information on most influential papers :

- Scopus term analyzer [6] for a time line for catch phrase "medical concept representation" used in titles of publications
- Ultimate Research Assistant [7] to extract contextual environment
- Wordle tag cloud tool to visualize the context [8]
- Identification of authors of **ten most influential papers** by using seven sources: *1) Web of Science, 2) Scopus, 3)Embase, 4) PubMed, 5) Google Scholar, 6) Cochrane Library, 7) British Library on-line catalogue.*
- Boolean search expression ***"concept representation" AND (medical OR medicine) AND (knowledge OR information)*** to filter non-relevant results
- two periods, 1988 – 1999 and 2000 – 2012, were compared by text-mining titles of publications

Setting up a social network survey:

- Primary source: LinkedIn group of the IMIA Medical Concept Representation WG, which has over 50 members with widely various backgrounds.
- Survey open from August 2012 to the end of October 2012.
- Secondary sources were additional LinkedIn Groups in broader domain where the survey was also published.
- Lists of most relevant papers shared by Datagle [9] and a tailored Google Docs document.

Reconciling bibliometrics and the survey data:

- noun phrase frequencies compared using text mining results of paper titles
- text mining performed with Textalyser [10]

Results 1

Figure 1- time line analytics results for the phrase "medical concept representation": the use of the catch phrase in the nineties and in the first decade of the new century. The timeline reveals that studies using this title phrase were done in mostly the nineties.

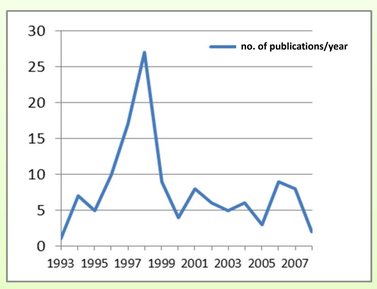
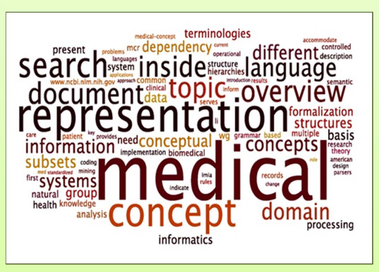


Figure 2- Tag cloud generated from result of catch phrase search reveals the wide contextual domain related to the catch phrase "medical concept representation".



Results 2

Table 1 – Set of thirty most influential authors of the period 1988-1999

Ranks were calculated by adding positional scoring of papers in the various lists, weighted in favour of multiple appearances of authors.

score	authors (1-15)	score	authors (16-30)
83.2	Cimino JJ	22.8	**Rosse C,**
43.2	Oliver DE	21	Miller R.A.
39.2	Baud RH	21	Rassinoux AM
39.2	Scherrer JR	19.2	**Musen MA**
35	Rector AL	19.2	Nowlan WA
33.6	Bell DS	18.2	Wagner JC
33.6	Shahar Y	18	Bailey KR
33.6	Shortliffe EH	18	Bauer BA
30.8	Fieschi M	18	Elkin PL
30.8	Huff SM	16.8	Chute CG
30.8	Joubert M.	15.6	Schoolman HM
30.8	Volot F.	15.6	Barnett GO
26.6	Johnson SB	15.6	**Horrocks I**
22.8	Evans DA	15.6	Humphreys BL
22.8	Hersh WR	15.6	Lindberg DAB

Table 2 – Set of most cited authors, period 2000 – 2012, in the **composite contextual domain** of authors on medical concept representation. (The full citation data was compared, not limited to citing the exact search phrase.) By comparing author's citation data over the composite contextual domain a broader coverage and less bias toward particular phrasing is achieved. Only about ten percent of mostly cited authors of the nineties remained in the list of the recent period.

authors (1-15)	cited	authors (16-30)	cited
Smith B	125229	Noy NF	21195
Roberts A	62871	Nadeau SE	20886
Stevens R	62715	Joffe H	19204
Horrocks I	60177	Wroe C	17530
Van Harmelen F	58626	Lussier, Y	14802
Fensel D	58491	Coronado S	14105
Zadeh LA	49420	Saraceno C	6976
Goble C	46984	Sioutos N	6584
Heilman KM	45858	Yao YY	5516
Decker S	41092	Shagina L	5407
Friedman C	40473	Hartel FW	3211
Pal SK	31200	Mejino JR	2975
Musen MA	28181	Haber MW	2325
Aspden, P	23482	Shiu SCK	1611
Rosse C	22253	Steinman F	1074

Results 3

Table 3 - A word frequency analysis of the titles of all `*medical concept representation*' papers of the period 1988-1999 shows the ten most frequently used single noun phrases and meaningful two-word phrases.

rank	single terms	two-word phrases
1	knowledge	medical language
2	language	natural language
3	concept	case based
4	clinical	knowledge representation
5	terminology	knowledge acquisition
6	data	language processing
7	representation	medical concept
8	information	medical terminology
9	model	structured data
10	system	concept representation

Table 4 – List of most frequently used single noun phrases and two word phrases of the period 2000-2012 in the domain of medical concept representation. All red entries are new terms or phrases in the titles of papers published, compared to the list of phrases of the 1988-1999 period.

rank	single terms	two-word phrases
1	health	health informatics
2	**information**	electronic health
3	**clinical**	**natural language**
4	**knowledge**	concept based
5	ontology	decision support
6	**case**	**language processing**
7	**data**	**concept representation**
8	semantic(s)	**medical language**
9	**concept**	medical informatics
10	**representation**	description logic

Results 4

Table 5 - title term changes : outgoing terms are in white. New title terms are red. The contextual domain of formal medical concept representation was broadened. Prominence of terms "semantics" and "ontology" shows a new paradigm. Terms "language", "model" and "terminology" disappeared - these more differentiated areas branched off from the previously common roots.

frequency rank	terms 1988-1999	terms 2000-2012
1	knowledge	health
2	language	information
3	concept	clinical
4	clinical	knowledge
5	terminology	ontology
6	data	case
7	representation	data
8	information	semantic(s)
9	model	concept
10	system	representation
below 10		

periods

Results 5

Table 6 - Results of the survey of socially active researchers got over 40 responses. The idea of using ontologies has clearly become the central paradigm. The list of top 20 titles of multiple occurrences are presented. Major research areas and resources in these lists include the **OBO** Foundry, the **Gene Ontology**, various versions of Systematized Nomenclature of Medicine (**SNOMED**) and the Unified Medical Language System (**UMLS**).

↓ Rank	Title ↓
1	The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration (*849*) [12]
2	The Unified Medical Language System (UMLS): integrating biomedical terminology (*709*) [13]
2	A reference ontology for biomedical informatics: the Foundational Model of Anatomy (*705*) [14]
2	Relations in biomedical ontologies (*672*) [15]
2	Desiderata for controlled medical vocabularies in the twenty-first century (*457*) [16]
2	Clinical terminology: why is it so hard? (*234*) [17]
2	From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies (*81*) [18]
2	BioCaster: detecting public health rumors with a Web-based text mining system (*63*) [19]
2	An ontology of epidemiological terms (ref website) [20]
3	Gene ontology: tool for the unification of biology (*10008*) [21]
3	Sweetening Ontologies with DOLCE (*668*) [22]
3	The medical dictionary for regulatory activities (MedDRA) (*198*) [23]
3	SNOMED clinical terms: overview of the development process and project status (*150*) [24]
3	Towards a reference terminology for ontology research and development i+n the biomedical domain (*102*) [25]
3	Methods in biomedical ontology (*92*) [26]
3	Ontology-based error detection in SNOMED-CT (*82*) [27]
3	Fuzzy Health, Illness, and Disease (*60*) [28]
3	Modeling biomedical experimental processes with OBI(*58*) [29]
3	Bringing epidemiology into the Semantic Web (*1*) [30]
3	A Dictionary of Epidemiology (book) [31]

* Frequency ranking is based on incidence in survey lists. The first ranked paper was mentioned the most times in various lists. Papers with rank '2' shared the second highest number of occurrence and so on. Papers with same rank are in order of their citation frequency, shown above in *(italics)*

Discussion and conclusions

Methodology issues:

- this is an ad hoc methodology not aiming at a new scientometric index or a generalized, reusable tool to measure or assess other areas
- combining on-line library databases, bibliometric services and text mining tools resulted in study-focused tool sets while large size of screened sources alleviate possible bias
- using the notion "concept" decreased probably due to the following factors: (A) propagation of the paradigm of ontological realism and critique of "conceptualism"[33], (B) substitution of the notion "concept" by the notion "class" in the Semantic Web and description logics community [34], (C) by the obvious ambiguity of the word itself [35]

Conclusions:

- the new millennium has coincided with a change in the focus of research
- the new paradigm of Semantic Web and results of ontology theory and practice have become new anchors [36].
- the central role of the term "concept" has been gradually abandoned
- the new paradigm could be illustrated with the following exemplars:
- **capture of medical information and knowledge leverages (standard based) ontologies**
- **open reference resources for content are developed collaboratively and are increasingly shared, reused**
- **web enabled standards help achieve transparent results.**

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

Reference list: http://goo.gl/EJZrg **Mail to: laszlo.balkanyi@ecdc.europa.eu**