

# From Concept Representations to Ontologies: A Paradigm Shift in Health Informatics?

Stefan Schulz, MD<sup>1</sup>, Laszlo Balkanyi, PhD<sup>2</sup>, Ronald Cornet, PhD<sup>3,4</sup>, Olivier Bodenreider, PhD<sup>5</sup>

<sup>1</sup>Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria; <sup>2</sup>European Centre for Disease Prevention and Control, Stockholm, Sweden; <sup>3</sup>Department of Medical Informatics, Academic Medical Center, Amsterdam, The Netherlands; <sup>4</sup>Department of Biomedical Engineering, Linköping University, Linköping, Sweden; <sup>5</sup>National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

**Objectives:** This work aims at uncovering challenges in biomedical knowledge representation research by providing an understanding of what was historically called “medical concept representation” and used as the name for a working group of the International Medical Informatics Association. **Methods:** Bibliometrics, text mining, and a social media survey compare the research done in this area between two periods, before and after 2000. **Results:** Both the opinion of socially active groups of researchers and the interpretation of bibliometric data since 1988 suggest that the focus of research has moved from “medical concept representation” to “medical ontologies”. **Conclusions:** It remains debatable whether the observed change amounts to a paradigm shift or whether it simply reflects changes in naming, following the natural evolution of ontology research and engineering activities in the 1990s. The availability of powerful tools to handle ontologies devoted to certain areas of biomedicine has not resulted in a large-scale breakthrough beyond advances in basic research.

**Keywords:** Data Mining, Terminology, Semantics, Vocabulary, Publishing

## I. Introduction

The study of the meaning of language expressions has a long history in health informatics, both regarding narratives (e.g.,

text in clinical reports and from the biomedical literature) and structured information (e.g., terms from standard vocabularies used for clinical research, health statistics, quality assessment and billing). It motivated the activities of the International Medical Informatics Association (IMIA)'s Working Group on Medical Concept Representation (MCR WG) [1], which was an influential body in the late 1980s and the 1990s, publishing regular overviews [2].

The evolution of ontologies for biomedical research, the proliferation of clinical vocabularies, advances in human language technologies with increasingly large amounts of training data have changed the health information science landscape profoundly. New scientific communities have arisen like the Semantic Web community, and social media are changing communication between researchers. In this context the MCR WG, now renamed to “Language and Meaning in Biomedicine (LaMB)”, will have to find a new ecological

**Submitted:** November 16, 2013

**Revised:** December 30, 2013

**Accepted:** December 30, 2013

### Corresponding Author

Stefan Schulz, MD

Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria. Tel: +43-316-385-13201, Fax: +43-316-385-13590, E-mail: stefan.schulz@medunigraz.at

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2013 The Korean Society of Medical Informatics

niche. In order to better define the future activities of this working group, the authors have investigated the evolution of the field of biomedical language and representation of meaning over the years, and will discuss some persistent research areas to be addressed in the future.

## II. Methods

The analysis of literature over time can provide insight in how a research field develops [3]. We have used bibliographics, on-line text mining tools and a social media survey tool, in order to investigate how the research area, known as “Medical Knowledge Representation” has evolved since the 1990s.

The phrase “medical concept representation” (not to be mixed with “concept representation” as a category used in the science of psychology) was key in that period—a reason to name the working group accordingly. Therefore, we placed this phrase in the centre of our investigation, divided into the following steps:

- Time line analysis of the occurrence of the phrase “medical concept representation” using the *Scopus* term analyser [4], extraction of the contextual environment using *Ultimate Research Assistant* [5] and visualization of the results using a tag cloud [6];
- Using the tool *Publish or Perish* [7] to identify the authors of the most influential papers, using seven sources, viz. *Web of Science*, *Scopus*, *Embase*, *PubMed*, *Google Scholar*, *Cochrane Library*, *British Library on-line catalogue*. The question was to have an idea of the persistence of the influential authors from the first period to the second one. The Boolean search expression “*concept representation*” AND (“*medical*” OR “*medicine*”) AND (“*knowledge*” OR “*information*”) was submitted to all of them, with variations according to their proprietary syntax. For identifying the top ten papers, the results of the seven lists were consolidated into a common table. For this, available citation ranks were taken, otherwise the source’s own ranking mechanism was used. In the following, the top ten papers were the source for extracting the top thirty authors, which were ranked in a second step. For this, the following heuristics was used: The  $n^{\text{th}}$  author in the list was assigned a score of  $11 - n$ , the eleventh and following authors was given a zero value. The scoring was weighted, favouring multiple appearances of authors in different sources: a final score was calculated as a net score ( $0.8 + 0.2 \times \text{occurrence}$ ).
- In the post-2000 analysis, due to the significant drop of the usage of the exact phrase “medical concept representation” the resulting paper population would have been too small for applying the same procedure as described

for the first period. Therefore, instead of summing up the citation data only for papers matching the query, here the citation data for all papers per author were used. This same method, however, could not be used for same analysis backwards to the previous period, due to limitations of the tool used [7].

- The hypothesis of a paradigm shift was studied, comparing relevant papers published during the years from 1988 to 1999 with those appearing between 2000 and 2012, focusing the same subject area. The reason for starting with 1988 was the availability of bibliographic databases, being almost accordant with the period of our interest, viz. the activities of the IMIA WG on *Medical Knowledge Representation*. Author lists were compared and all the titles of the two full paper sets were text mined using *Textalyser* [8].
- The second, more recent set was cross-checked against a third set from the same period, obtained by an online survey targeted to the specifically interested audience. For this survey (open from August to October 2012) the primary source was the *LinkedIn* group of the MCR WG, having at that time over fifty members of widely various backgrounds. Secondary sources were additional *LinkedIn* Groups in broader domain. Participants were asked to quote and to share the papers they found to be most influential in their work or research. We used *Datagle* [9] and a Google document to collect survey data.

## III. Results

### 1. Looking Back: ‘Medical Concept Representation’ before the Turn of the Millennium

Scopus has revealed that the exact phrase “medical concept

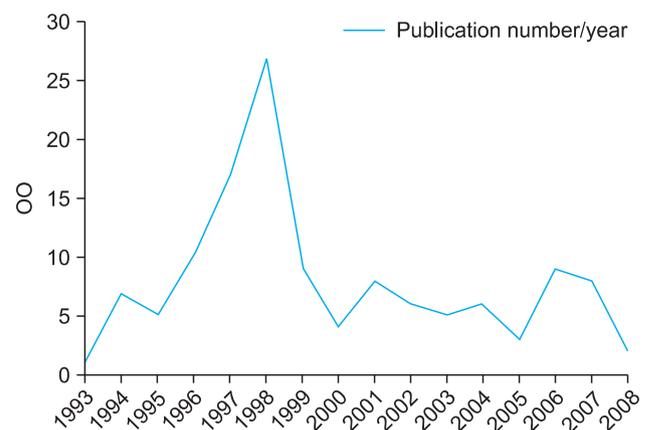


Figure 1. Scopus time line analytics results for the exact phrase “medical concept representation”.



Table 2. List of most frequently used uni- and bi-grams of the period 1988–1999

Rank	Words	Word bi-grams
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 3. Set of most cited authors, between 2000 and 2012, covering the whole domain of all authors publishing on medical concept representation

Authors (1–15)	Cited	Authors (16–30)	Cited
Smith B	125,229	Noy NF	21,195
Roberts A	62,871	Nadeau SE	20,886
Stevens R	62,715	Joffe H	19,204
<b>Horrocks I</b>	60,177	Wroe C	17,530
Van Harmelen F	58,626	Lussier, Y	14,802
Fensel D	58,491	Coronado S	14,105
Zadeh LA	49,420	Saraceno C	6,976
Goble C	46,984	Sioutos N	6,584
Heilman KM	45,858	Yao YY	5,516
Decker S	41,092	Shagina L	5,407
Friedman C	40,473	Hartel FW	3,211
Pal SK	31,200	Mejino JR	2,975
<b>Musen MA</b>	28,181	Haber MW	2,325
Aspden P	23,482	Shiu SCK	1,611
<b>Rosse C</b>	22,253	Steinman F	1,074

Names that are in the 1988–1999 ranking are in bold face.

representation domain. The fact that “language”, “model” and “terminology” disappeared may suggest that some more differentiated areas branched off the previously common roots.

#### 4. Results of the Survey Taken Show the Opinion of Socially Active Researchers Interested in the Domain

The survey had 42 respondents. Not surprisingly, the central role of ontologies is clearly reflected in the list of the twenty most influential papers (Table 5). Recurring resources in-

Table 4. List of most frequently used uni- and bi-grams of the period of the period 2000–2012 in the domain of medical concept representation

Rank	Words	Word bi-grams
1	health	health informatics
2	<b>information</b>	electronic health
3	<b>clinical</b>	<b>natural language</b>
4	<b>knowledge</b>	concept based
5	ontology	decision support
6	<b>case</b>	<b>language processing</b>
7	<b>data</b>	<b>concept representation</b>
8	semantic(s)	<b>medical language</b>
9	<b>concept</b>	medical informatics
10	<b>representation</b>	description logic

Bold face highlights the terms that also occur in the top-ten list from the 1988–1999 period (Table 2).

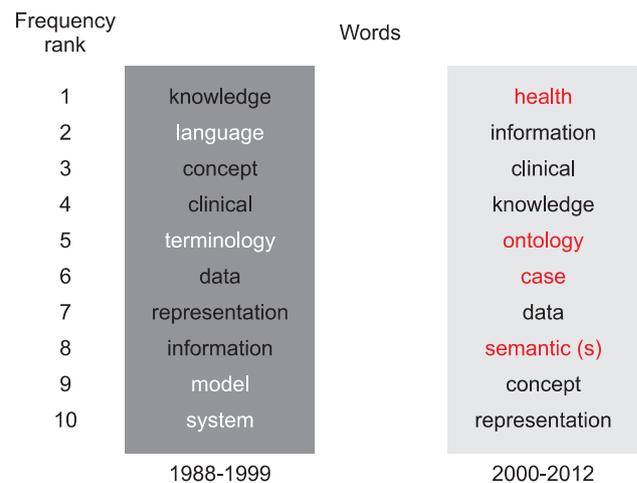


Figure 3. Changes in the most frequent title words of papers on medical concept representation.

clude the Open Biological and Biomedical Ontologies (OBO) Foundry [10], the Gene Ontology [11], Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [12], and the Unified Medical Language System (UMLS) [13].

## IV. Discussion

### 1. Methodology Issues Regarding the Literature Study

Although the methodology applied in this paper does not aim at establishing a new scientometric index or a generalizable tool, it clearly demonstrated that on-line searchable library databases, bibliometric services, and simple text mining tools enable the creation of study-focused tool sets as used in this study without investing much effort and re-

Table 5. Titles of the twenty most influential papers as listed by LinkedIn MCR WG members

Rank <sup>a</sup>	Title
1	The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration (849) [10]
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]
3	Gene ontology: tool for the unification of biology (10,008) [11]
3	Sweetening ontologies with DOLCE (668) [20]
3	The medical dictionary for regulatory activities (MedDRA) (198) [21]
3	SNOMED clinical terms: overview of the development process and project status (150) [22]
3	Towards a reference terminology for ontology research and development in the biomedical domain (102) [23]
3	Methods in biomedical ontology (92) [24]
3	Ontology-based error detection in SNOMED CT (82) [25]
3	Fuzzy health, illness, and disease (60) [26]
3	Modeling biomedical experimental processes with OBI (58) [27]
3	Bringing epidemiology into the Semantic Web (1) [28]
3	A dictionary of epidemiology (book) [29]
3	Semantic interoperability for better health and safer healthcare [30]

MCR WG: Working Group on Medical Concept Representation, OBO: Open Biological and Biomedical Ontologies, SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms, OBI: Ontology for Biomedical Investigations.

<sup>a</sup>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*.

sources. Using multiple, large bibliographic source databases helped to alleviate the possible bias in such studies that are limited to one particular source or aspect of the field.

## 2. Current Trends

The tools we used in this study were aimed at exploring the specific area of medical concept representation with the focus on testing the complementary question as to whether the observed changes amount to a significant paradigm shift.

Our results show that researchers active in this area for several decades have pursued the main goal of being able to make health-related information machine readable and processable. This has been a major driver of the development of clinical information systems in general. The use of formal languages, such as description logics, has been a step in this direction. In 1990s, “medical concept representation” was seen as a solution by proposing just one general method: practical conceptualization of information in medical re-

search and practice. However, these efforts were hindered by theoretical issues, difficulties of modelling a domain, and the explosion of knowledge in general [31].

Building on this background, our investigation has taken the pulse of a group of researchers interested in what we could refer to, generally speaking, as the study of meaning of structured and unstructured representations. First of all the use of the term “concept” has decreased, which we attribute to the following factors:

- Propagation of the paradigm of ontological realism, the proponents of which have been arguing against the usage of this word in the context of ontologies, contending that the representation of concepts as “entities of thought” is inappropriate for the representation of a scientific domain and obfuscates the difference between the entities and names given to them [32];
- The preference of “class” over “concept” in the Semantic Web and description logics community, especially regard-

ing the influential OWL family of representation languages [33];

- The obvious polysemy of the word itself [34].

In addition, the popularity of the word “ontology” shows a new tendency in which artefacts that represent types of domain entities are more clearly distinguished by some researchers from artefacts that describe language items. The importance of ontology-based artefacts can be seen by the central place the OBO Foundry and SNOMED CT occupy in publications and importance judgments. However, the boundaries between ontologies and knowledge representation artefacts are less clear, although relatively crisp criteria can be formulated. In practice, “ontology” is used by many to refer to a wide array of resources across the semantic spectrum, encompassing terminologies, thesauri, classifications and formal ontologies [35].

At the same time important areas as medical language processing and medical terminologies, but also metadata, semantic annotation and folksonomies have gained importance, so that they are no longer subsumed under “concept representation”.

The analysis of influential authors faced methodological difficulties, as the selection criterion—namely the phrase “concept representation” turned out to be a moving target. The comparability of the two lists of authors is therefore limited. Nevertheless, it is noteworthy that only three authors appeared on both lists. Note that this comparison is additionally biased by the following: it is very likely that there are relevant authors in the second period that were not retrieved, simply because they did not use the—already outmoded—phrase “concept representation”, at all. There are authors of the papers in Table 5 that are not among the top 20 (Table 4), simply because they avoid that phrase. If they would have been included, the overlap were probably even lower.

## V. Conclusion

There are several indications that the turn of the new millennium coincided with a change in the focus of research in medical domain representation and semantics. The millennium marked the emergence of the establishment of applied ontology [36] and the Semantic Web [37] as new disciplines. The central role of the term “concept” has been gradually abandoned. Whether this really amounts to a paradigm shift, or a simple change in terminological preferences, may be argued. Undoubtedly, the ontology research and engineering efforts, which started around 1990, yielded important results, including the development of description logics [38], tools like Protégé [39], as well as the groundbreaking GA-

LEN project [40].

The following directions for the future have emerged from our analysis:

- The capture of medical information and knowledge leverages (standards) ontologies;
- Open reference resources for content are developed collaboratively, shared, and reused;
- Web enabled standards help achieve transparent results;
- “Big data” opens new ways for knowledge acquisition;
- However, a large part of clinical information continues being recorded as free text, which keeps the need of processing medical language on the research agenda.

All these topics justify, more than ever, collaborative research and development efforts, for which the IMIA WG Language and Meaning in Biomedicine (LaMB) [41] can be an effective catalyst.

## Conflict of Interest

No potential conflict of interest relevant to this article was reported.

## Acknowledgments

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. We also thank participants of the IMIA LaMB Working Group for their participation in the survey.

## References

1. International Medical Informatics Association [Internet]. Geneva, Switzerland: International Medical Informatics Association; c2013 [cited at 2013 Nov 15]. Available from: <http://www.imia-medinfo.org/new2/>.
2. Cimino JJ, Smith B. Introduction: international medical informatics association working group 6 and the 2005 Rome conference. *J Biomed Inform* 2006;39(3):249-51.
3. Schuemie MJ, Talmon JL, Moorman PW, Kors JA. Mapping the domain of medical informatics. *Methods Inf Med* 2009;48(1):76-83.
4. Scopus [Internet]. Amsterdam, The Netherlands: Elsevier; c2013 [cited at 2013 Nov 15]. Available from: <http://www.scopus.com>.
5. Ultimate Research Assistant [Internet]. Herndon (VA): Andy Hoskinson, LLC; [cited at 2013 Nov 15]. Available from: <http://ultimate-research-assistant.com>.
6. Wordle [Internet]. [unknown]: Jonathan Feinberg; c2013 [cited at 2013 Nov 15]. Available from: <http://>

- www.wordle.net/.
7. Harzing AW. Publish or Perish [Internet]. [unknown]: Harzing.com; c2013 [cited at 2013 Nov 15]. Available from: <http://www.harzing.com/pop.htm>.
  8. Textalyzer [Internet]. [unknown]: textalyzer.net; c2004 [cited at 2013 Nov 15]. Available from: <http://textalyser.net/>.
  9. Datagle [Internet]. [unknown]; Datagle LLC; [cited at 2013 Nov 15]. Available from: <http://www.datagle.com/>.
  10. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25(11):1251-5.
  11. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25(1):25-9.
  12. International Health Terminology Standards Development Organisation. SNOMED CT [Internet]. Copenhagen, Denmark: International Health Terminology Standards Development Organisation; c2013 [cited at 2013 Nov 15]. Available from: <http://www.ihtsdo.org/snomed-ct>.
  13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(Database issue):D267-70.
  14. Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform* 2003;36(6):478-500.
  15. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. *Genome Biol* 2005;6(5):R46.
  16. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med* 1998;37(4-5):394-403.
  17. Rector AL. Clinical terminology: why is it so hard? *Methods Inf Med* 1999;38(4-5):239-52.
  18. Smith B. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. *J Biomed Inform* 2006;39(3):288-98.
  19. Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, et al. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics* 2008;24(24):2940-1.
  20. Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L. Sweetening ontologies with DOLCE. In: *Knowledge engineering and knowledge management: ontologies and the Semantic Web*. Heidelberg, Germany: Springer; 2002. p. 166-81.
  21. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf* 1999;20(2):109-17.
  22. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. *Proc AMIA Symp* 2001;2001:662-6.
  23. Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a reference terminology for ontology research and development in the biomedical domain. In: *Proceedings of KR-MED; 2006 Nov 8; Baltimore, MD*. p. 56-67.
  24. Yu AC. Methods in biomedical ontology. *J Biomed Inform* 2006;39(3):252-66.
  25. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. *Stud Health Technol Inform* 2004;107(Pt 1):482-6.
  26. Sadegh-Zadeh K. Fuzzy health, illness, and disease. *J Med Philos* 2000;25(5):605-38.
  27. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, et al. Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 2010;1 Suppl 1:S7.
  28. Ferreira JD, Pesquita C, Couto FM, Silva MJ. Bringing epidemiology into the Semantic Web. In: *Proceedings of the 3rd International Conference on Biomedical Ontology; 2012 Jul 21-25; Graz, Austria*.
  29. Porta M. *A dictionary of epidemiology*. 5th ed. New York (NY): Oxford University Press; 2008.
  30. European Commission. *Semantic interoperability for better health and safer healthcare: research and development roadmap for Europe*. Luxembourg: European Commission; 2009.
  31. Gillam M, Feied C, Handler J, Moody E, Shneiderman B, Plaisant C, et al. The healthcare singularity and the age of semantic medicine. In: Hey T, Tansley S, Tolle K, editors. *The fourth paradigm: data-intensive scientific discovery*. Redmond (WA): Microsoft Research; 2009. p. 57-64.
  32. Smith B. Beyond concepts: ontology as reality representation. In: *Proceedings of the 3rd International Conference on Formal Ontology in Information Systems; 2004 Nov 4-6; Torino, Italy*. p. 73-84.
  33. W3C OWL Working Group. *OWL 2 Web ontology language document overview (second edition)* [Internet]. [unknown]: W3C; c2012 [cited at 2013 Nov 15]. Available from: <http://www.w3.org/TR/owl2-overview/>.
  34. Klein GO, Smith B. Concept systems and ontologies: recommendations for basic terminology. *Trans Jpn Soc Artif Intell* 2010;25(3):433-41.
  35. Schulz S, Jansen L. Formal ontologies in biomedical knowledge representation. *Yearb Med Inform*

- 2013;8(1):132-46.
36. Smith B. Applied ontology: a new discipline is born. *Philos Today* 1998;12(29):5-6.
  37. Berners-Lee T, Hendler J, Lassila O. The Semantic Web. *Sci Am* 2001;284(5):28-37.
  38. Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider P. *The description logic handbook: theory, implementation, and applications*. 2nd ed. Cambridge, UK: Cambridge University Press; 2007.
  39. Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubezy M, Eriksson H, et al. The evolution of Protégé: an environment for knowledge-based systems development. *Int J Hum Comput Stud* 2003;58(1):89-123.
  40. Rector AL, Glowinski AJ, Nowlan WA, Rossi-Mori A. Medical-concept models and medical records: an approach based on GALEN and PEN&PAD. *J Am Med Inform Assoc* 1995;2(1):19-35.
  41. IMIA LaMB Working Group [Internet]. Mountain View (CA): LinkedIn; c2013 [cited at 2013 Nov 15]. Available from: <http://www.linkedin.com/groups/IMIA-Medical-Concept-Representation-Working-3680642/about>.