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

Stefan Schulz, Laszlo Balkanyi, Ronald Cornet and Olivier Bodenreider

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:
http://dx.doi.org/10.4258/hir.2013.19.4.235  
Copyright: © 2013 The Korean Society of Medical Informatics  
http://synapse.koreamed.org/  
Postprint available at: Linköping University Electronic Press  
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-107073
From Concept Representations to Ontologies: A Paradigm Shift in Health Informatics?

Stefan Schulz, MD¹, Laszlo Balkanyi, PhD², Ronald Cornet, PhD¹,4, Olivier Bodenreider, PhD⁵

¹Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria; ²European Centre for Disease Prevention and Control, Stockholm, Sweden; ³Department of Medical Informatics, Academic Medical Center, Amsterdam, The Netherlands; ⁴Department of Biomedical Engineering, Linköping University, Linköping, Sweden; ⁵National 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...
niches. 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 into 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^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 × 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 representation”...
representation” was used mostly in the nineties (Figure 1). Scopus data were available for 1993–2008. The targeted semantic search revealed a wide conceptual domain related to this phrase, as shown in Figure 2.

The top thirty authors of the ten most influential papers 1988–1999 were identified (the starting date of the study was justified by the availability of electronic bibliographic databases and the comparability of the investigated periods before and after 2000). The tool Publish or Perish [7] showed the average number of authors to be 2.45. The results of the extraction of the first three author names per paper are shown in Table 1. Our querying strategy was found effective for excluding papers regarded irrelevant for our purpose, e.g., in the domain of concept representation in psychology.

A frequency analysis of the title words of the papers in the same period shows the most frequently used uni- and bi-grams (single noun phrases and meaningful two-word phrases) in Table 2. Note that ‘ontology’ was not among the most frequently used terms at that period.

2. “Medical Concept Representation” Since the Year 2000

Table 3 presents the list of the top thirty authors of most cited publications, using the same Boolean expression applied to the period of 2000–2012. However, as the methodology was different for the reasons explained above, the comparison should be interpreted with reservation. Nevertheless it is striking that the two lists only overlap in three authors (in bold). In addition, the word frequency analysis of the period 2000–2012 shows a clearly distinct result (Table 4).

3. Mapping the Conceptual Context of Most Influential Papers Based on Text Mining of Titles

Figure 3 shows how the terms in the titles were changing. “Old” terms that are no longer found among the “new” top ten are depicted in white. New terms appearing in the 2000–2012 list are shown in red. The top ten terms also suggest that the subject matter of “concept representation” was broadened (from focusing on “medical” to areas as “health” and “clinical”). In addition, the words “semantics” and “ontology” suggest that new ideas have influenced the concept...
representation domain. The fact that "language", "model" and "terminology" dissolved 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-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 2. List of most frequently used uni- and bi-grams of the period 1988–1999

<table>
<thead>
<tr>
<th>Rank</th>
<th>Words</th>
<th>Word bi-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>knowledge</td>
<td>medical language</td>
</tr>
<tr>
<td>2</td>
<td>language</td>
<td>natural language</td>
</tr>
<tr>
<td>3</td>
<td>concept</td>
<td>case based</td>
</tr>
<tr>
<td>4</td>
<td>clinical</td>
<td>knowledge representation</td>
</tr>
<tr>
<td>5</td>
<td>terminology</td>
<td>knowledge acquisition</td>
</tr>
<tr>
<td>6</td>
<td>data</td>
<td>language processing</td>
</tr>
<tr>
<td>7</td>
<td>representation</td>
<td>medical concept</td>
</tr>
<tr>
<td>8</td>
<td>information</td>
<td>medical terminology</td>
</tr>
<tr>
<td>9</td>
<td>model</td>
<td>structured data</td>
</tr>
<tr>
<td>10</td>
<td>system</td>
<td>concept representation</td>
</tr>
</tbody>
</table>

Table 3. Set of most cited authors, between 2000 and 2012, covering the whole domain of all authors publishing on medical concept representation

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

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

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

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

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


35. Schulz S, Jansen L. Formal ontologies in biomedical knowledge representation. Yearb Med Inform