to the clusters to identify key concepts. In this way, we let the underlying structure of the literature speak for itself. It would be interesting to see how these concepts could be related to the HIO.

12. Comment by B. Smith

It Usually Begins with the Gene Ontology

Biomedical ontologies have now become a standard part of the biomedical informatician's toolkit. Initially, with the Gene Ontology (http://geneontology.org), ontologies were introduced to enhance the comparability of gene array data deriving from research on different model organisms. Very rapidly they began to be used as tools to enhance the discoverability of data more generally, to allow new sorts of statistical analyses of data under the heading of 'gene enrichment studies' and to allow the merger of large bodies of data deriving from different sources through the use of a common set of ontology annotations.

This latter application is of increasing importance especially in translational medicine and in interdisciplinary areas such as research on aging, where ontologies are playing what we might think of as an educational role. In aging research, for example, researchers working on nutrient sensing might be called upon to collaborate with those working on mitochondrial dysfunction, or on stem cell exhaustion, and all of these might in turn need to collaborate with experimentalists working on apoptosis in yeast. To a surprising degree the Gene Ontology is serving as a common resource upon which all of these communities are able to draw in combining their data. I believe that part of what is going on here is that, when human beings need to formulate and test hypotheses and to display and analyze experimental results involving contributions from unfamiliar disciplines, the GO is used, in effect, as a simple educational aid.

But the FMA Is also Involved

In fact researchers in biomedical ontology already from the very start have been suggesting that ontologies might serve an educational role of an even more ambitious sort. Cornelius Rosse and his collaborators in Seattle proposed as early as 1998 that ontologies could serve as a platform to re-engineer education in the core basic sciences. The discipline of anatomy, as they pointed out

is the first subject – and one of the most challenging and time-consuming subjects – introduced in the training of all health care professionals. There is a need for logic-based, machine-parsed representations of anatomical knowledge for the creation of intelligent educational programs in anatomy.

What they at that stage referred to as the 'Digital Anatomist ontology' and has since been transformed into the Foundational Model of Anatomy Ontology [45, 46] would, they held, establish 'a basic requirement for such applications' and would 'serve as a platform for a digitally re-imagined approach to the teaching of anatomy as core basic science in medical education programs' [47].

This 'digitally re-imagined approach' would then be applied not merely through the FMA anatomy reference ontology but also through reference ontologies in other areas of basic science, including genetics, cell biology, physiology, and so forth. A single set of reference ontologies would in this way – given the full realization of Rosse's vision – become engrained in the course of medical training on the very brains of medical students. These ontologies would then automatically work in tandem with the ontologies being used to capture the clinical data which these medical students are using in their daily activities, since the latter would be built up on the basis of the former.

We Are What We Publish

Elkin, Brown and Wright, in their "Biomedical Informatics: We Are What We Publish" [1], formulate what we can think of as an ambitious complement to Rosse's vision. They argue in effect that we can not merely use ontologies as a vital tool of medical education, but that we can go further and use the ontological approach to determine the very content of one (and not the least important) branch of the biomedical curriculum. They make this proposal in the context of an analysis of the AMIA and IMIA initiatives to formalize the definition of 'biomedical informatics,' extracting to this end the terms used in the AMIA consensus document and combining these with the terms employed in the IMIA definitions. They then built manually on this basis a draft Health Informatics Ontology, which they used to parse a very large corpus of medical literature identified using NLP software, with "Medical Informatics OR Bioinformatics" as search criterion.

The results are of interest from a number of different points of view. But they show that the merged AMIA-IMIA-based ontology is able to identify the coverage domain of biomedical informatics only partially, in that of the 168,298 articles identified, only some 37% contained at least one term from the HIO in its title or abstract. Work is accordingsly on-going on a new version of the HIO, both expanded and more formal, in order to establish the degree to which there is material published in the field of biomedical informatics that is not covered by the AMIA/IMIA specifications.

Such an expanded HIO could then be used for more ambitious investigations – for example to provide a series of snapshots of the discipline to demonstrate how it has changed, and is still changing, over time. The enhanced ontology would contribute, as the authors point out, to a greater self-understanding of the discipline of biomedical informatics by its practitioners – and it could thereby also help to realize the vision for ontology as a tool for biomedical informatics education along the lines proposed by Rosse.

At the same time, however, we can see some of the problems facing such a vision. As the authors acknowledge, the HIO itself is still in early draft stage, and it lacks formal definitions of its constituent terms. It was moreover developed on the basis of inputs created through both an AMIA and an IMIA consensus process that was not aimed at yielding an ontological representation of a principled sort. The result requires work to adapt the HIO to best practice principles for ontology development, including those identified through the OBO Foundry initiative (http://obofoundry.org).

To be of value to the process of biomedical education, integration of HIO with the reference ontologies corresponding to the basic biomedical sciences would also be important. Building ontologies using what the authors call 'concepts' in the biomedical literature and relying on the HIO to provide semantic context along the lines the authors propose will yield satisfactory results only if the HIO itself is in good shape from an ontologico-semantic point of view, and for this considerable further effort is needed. The results should then satisfy not merely consensus review by the practitioners of the specialty of biomedical informatics, but also survive stringent examination by specialists in the field of ontology.

Creating an HIO in this manner will be no easy task. In contrast to human anatomy, which is an evolutionarily highly stable domain marked by a considerable degree of disciplinary self-understanding, biomedical informatics is an inherently complex and interdisciplinary and above all dynamically evolving field. As the GO has shown, an ontology can demonstrate considerable practical value even in a rapidly changing field of scientific endeavor. Having taken it upon themselves to create the Health Informatics Ontology, the authors now have the responsibility to work with the ontology community to demonstrate that they can together create an artifact marked by the sort of ontological rigor that would make it truly useful in defining and shaping the field of biomedical informatics.

13. Comment by J. Talmon

The title of the Elkin et al. paper [1] raises high expectations. One would expect to find a description of our field based on what we have published. At the end of the manuscript we are still (partially) in the dark.

As a frame of reference, the authors used the Health Informatics Ontology (HIO). Their main finding is that only 37% of the retrieved articles had a concept in the title or abstract that also occurred in the HIO. On the other hand only 251 of the 433 concepts of the HIO could be identified in the literature. One would have expected that a deeper analysis was made of the articles that did not have a concept of the HIO as well as of the concepts in the HIO that were not found in the literature. Such an analysis would have revealed were the discrepancy is and whether or not we should revise the top-down developed AMIA and IMIA terminologies. At least some supplementary material should have been provided to allow the reader to better understand these discrepancies.

The analysis of the random sample of 27,000 articles with abstracts by parsing them with SNOMED-CT seems too much focused on disorders rather than clinical subspecialties. Unfortunately there is no data on the number of articles that did not have a SNOMED-CT code for a disorder by body site.

Apart from the disappointment related to the limited analyses and the rather general discussion, there are a few methodological issues I would like to raise.

A main concern is the method used to define what is being published in our field. In the Schuemie et al. paper [13] we started from what is being published in the journals that are defined by Thomson Reuters to cover the field of Medical Informatics; indeed we did not consider bioinformatics. From there we identified which 1-, 2-, and 3-grams were more common in the corpus of MI publications as compared to the rest of the PUBMED corpus. We also investigated whether there are other series or journals indexed in PUBMED that also published in our domain. As a matter of fact we tried to identify how our field differentiates itself from other disciplines, not on what we may have in common.

Elkin et al. on the other hand relied on two search terms to retrieve articles from PUBMED. One should be aware that searching *Medical Informatics* is different from searching for *"Medical Informatics"*. The former resulted in 153,403 hits, the latter, however, in only 9172 hits[d].This makes it clear that what you search for will influence the results, and thereby what our domain entails.

It is strange that only *Medical Informatics* and *Bioinformatics* have been used as search terms. In their paper, Elkin et al. also use the term biomedical informatics. This term gives 1815 hits in PUBMED, 260 of which were not covered by searches for *Medical Informatics* or *Bioinformatics*.

It seems that large areas of application of ICT in health care have not been covered. For example, Telemedicine – 8408 hits of which 5332 were not covered by the search for *Medical Informatics* or *Bioinformatics* – is hardly dealt with.

A further concern is that not all of what we publish in our Medical Informatics Journals has been retrieved by the queries of Elkin et al. IJMI, MIM, JAMIA and JBI have published 4103 articles prior to February 2006. Of those only 2871 appeared in queries for *medical informatics* or *bioinformatics*. Nearly one third of what is being published in our journals is not accounted for. Are all those papers outside the domain of (Bio)medical Informatics?

A final note is on the time span of the search used by Elkin et al. They performed their search in February 2006. We are now more than seven years later. In that time, the body of literature on (bio)medical Informatics has more than doubled (▶Table 1). In particular given the large increases in the number of publications in bio(medical)informatics, this raises the question how valid the findings still are.

Table 1 Number of publications at different time instances of search and their percentage increase

Query	Before February 2006	Before November 2013	Percentage increase
Medical Informatics	153,403	307,700	100%
Bioinformatics	32,238	125,655	290%
Telemedicine	8,408	16,254	93%
"our journals"	4,103	7,457	81%
Biomedical Informatics	1,815	6,492	257%

[d] All PUBMED searches have been done for publications prior to 01/02/2006.