Beyond the Tumour: Breast Cancer Phenotypes

—Towards a pluralistic integration of heterogeneous representations—

Aleksandra Sojic[1,2,3*] and Oliver Kutz[4]

[1]European School of Molecular Medicine, [2]European Institute of Oncology; Milan, Italy
[3]University of Milan, Milan, Italy
[4]Research Center on Spatial Cognition (SFB/TR 8), University of Bremen, Germany

ABSTRACT

Starting from an acknowledgment of the plurality of epistemic motivations driving phenotype representations, our main contribution is a distinction between six categories of *human agents as individuals and groups* focused around particular epistemic interests. We analyse the corresponding impact of these groups and individuals on representation types, mapping and reasoning scenarios, using the example of breast cancer research. We in particular demonstrate a heterogeneity of representation types for breast cancer phenotypes and stress that the characterisation of a tumour phenotype often includes parameters that go beyond the representation of a corresponding empirically observed tumour, thus reflecting significant functional features of the phenotypes as well as epistemic interests that drive the modes of representation. Accordingly, the represented features of cancer phenotypes function as epistemic vehicles aiding various classifications, explanations, and predictions.

1 INTRODUCTION

The representation of phenotypes plays an important role in clinical and biomedical knowledge. Besides functional characterisations, a disease often gets characterised through a distinction between 'normal' and 'abnormal' phenotypes, where 'abnormal' phenotypes often serve as the marks of disease. The 'abnormal' phenotypes associated with a disease are labelled as *phenotypes of disease* (PD). However, the questions of what is 'abnormal' and *what* should be considered as a phenotype of a disease and *how* such a phenotype should be represented are rather contentious. Clearly, the choice of how a PD should be represented is *normative* and *context dependent*. Consider the case of breast cancer and BRCA gene mutations. In the age of genomic medicine, the very definition of disease has changed introducing an asymptomatic diagnosis. So, carriers of BRCA mutation, without having developed any signs of breast cancer, still have a likelihood of over 80% for developing an aggressive cancer phenotype during their life span. Genomic medicine shifts the focus of PD from a traditional organ level approach to the gene level, treating apparently healthy people as 'patients'. For, the 'normal' breast phenotype in a BRCA mutation carrier will be irrelevant in the light of knowledge about 'abnormal', fine-grained phenotypes related to the gene expression patterns of the mutated gene. Although these new directions in biomedicine aim towards an integration of clinical and biomedical knowledge, in most cases the needs of sub-domain knowledge significantly vary. So, a clinician will have different criteria for a representation than a molecular biologist. Regarding the goals of a discipline and the research context, a representation that is relevant for a clinician does not need to satisfy the needs of a molecular biologist who is aiming towards more fine-grained representations. As a result, heterogeneous representations of breast cancer phenotypes were employed in clinical and biomedical knowledge [8, 4, 25].

Taking a very general position, representations of PDs may include images acquired by technologies such as ultrasound, X-ray, and microscopy of histopathological samples. Moreover, representations of PDs are not limited to visual representations, but may include mathematical equations, statistical graphs, molecular markers, microarrays data, and the phenotype specific protein interactions, thus describing PDs according to the needs of and knowledge about a particular domain aspect. In addition, a specific representation of a phenotype should not, in general, be mistaken for the representation of knowledge. Rather, a representation reflects which aspects of knowledge have been targeted by the representation. Accordingly, a representation reflects a scientist's choice of a representation type in order to represent a certain subset of the domain knowledge—therefore, 'choosing a representation' might be a highly intentional act [6]. However, a representation such as a histopathological image will not, itself,

*Corresponding author: aleksandra.sojic@ifom-ieo-campus.it

represent any knowledge unless it gets interpreted. Knowledge within a domain is explicitly represented only if the representations get systematically connected with related interpretations, knowledge claims, and reasoning over the representations. Therefore, besides heterogeneity of PDs, biomedical ontology has to deal with a heterogeneity of reasoning about PDs, comprising different kinds of formal (or logical) representations as well as various types of reasoning. Conversely, the intended reasoning methods or types over PDs also influence the choice of representation of PDs because such representations are mediated by domain specific methods and interventions, employed in the imaging, measuring of the gene expression and other diagnostic techniques [12]. For example, the clinical representation of breast cancer goes beyond the tumour imaging representation. According to the standards of the TNM classificatory system [8], the clinical classification of tumours might consider tumour size (T), lymph nodes involvement (N), and presence of metastasis (M). Of course tumour size is just one feature and is not sufficient for the characterisation of the tumour type. Cancer is a dynamic and complex disease of an organism and the PDs go beyond the characterisation of a tumour's features captured in a static picture. So, for example, knowledge about lymph nodes' status or proliferation marker KI-67 provides additional information about a tumour's phenotype. Likewise, tumour markers provide a view on the PDs through the specific interventions on the representation such as staining samples in order to mark the presence of hormone receptors. Had the estrogen receptor (ER) been detected, the PD would have been described as an ER positive tumour, which significantly differs from an ER− (negative) tumour, which does not respond to the endocrine therapy [7]. Thus, the therapeutic criteria are also considered in the specification of the tumour phenotypes.

2 A PLURALITY OF DOMAIN INTERESTS

Information technologies and formal tools such as ontologies for knowledge representation (KR) are aiming at the integration of heterogeneous knowledge domains and different types of representations. Concurrently, clinicians and molecular oncologists are trying to organise and apply the overwhelming and diverse knowledge about cancer biology. Can these interests of different disciplines meet in a constructive union, while preserving the domain specific representations and reasoning capabilities?

In this and the next section we outline some of the requirements for achieving such a level of interoperability.

We begin by giving a comparative analysis of the distribution and character of knowledge involved in the integration of heterogeneous types of knowledge represented in knowledge bases (KBs). In particular, we distinguish *where*, *how*, and *by whom* knowledge is represented by characterising six epistemic groups, and by discussing how membership to a group impacts the representation as well as knowledge base types. Note that these groups exhibit rich interdependencies and partially overlap.

1. The characterisation of the epistemic groups starts with the societal demands for problem solving, such as, for example, the need for personalised breast cancer therapy. The demands may be represented in the form of standards, platforms and funding policies. In a democratic society, knowledge on this level can be represented as common or shared knowledge available to the members of society; knowledge can be distributed through various channels or common-sense KBs.
2. The second epistemic group to be discussed is at the level of an individual scientist whose 'knowledge base' is a collection of relevant background knowledge, here to be understood as cognitive representations placed in the mind, arguably, in the form of conceptual maps (see [24]).
3. As the third epistemic group, we specify the scientific communities, each of which is composed of the specific disciplinary domain scientists (clinicians, molecular biologists, bioinformaticians etc.). This epistemic group establishes knowledge within a scientific community as a received view, having the form of *explicit* and *inter-subjective* representations expressed in the respective scientific languages, circulated through publications. Like in group (1), knowledge can be distributed in various ways, but related KBs will contain domain specific knowledge.
4. The fourth group comprises scientific communities formed around a particular problem (e.g. breast cancer). As the group contains multidisciplinary teams focused on a particular problem, knowledge will need to be coordinated in such a way that the used scientific terms and reference classes will conform with knowledge within diverse domains. For instance, the biomedical terms might be structured into networks of terms that represent how these terms are interrelated in the domain knowledge. Thus, collaboration here results in merging knowledge from different domains. The representation of the merged knowledge coming from different perspectives on the same problem

might be a 'unified semantic map' (see group (2)) that serves as a semi-formal conceptual model and an intermediate step towards the KB and the formal ontology to be employed in KR.

5. The fifth is the communities of logicians and ontologists who are formalising ontologies according to the needs and specificities of a particular field. Domain knowledge and the merged domain knowledge will be expressed as ontologies written in various formal languages (e.g. refining foundational ontologies such as DOLCE [21], BFO[1], or GFO[2] etc. formalised in OWL[3], first-order logic, etc.)
6. The sixth group involves computer scientists, programmers and engineers, who are designing databases and applying formal ontologies as well as various reasoning tools to large datasets. Technically, a representation built on top of a database involves types and mapping relations *structuring* the data, and can be considered as meta-data. Here the representation integrates the types and mappings with instances (data). Epistemic accuracy of the mappings depends on how well the mappings correspond to the scientific knowledge and the empirical findings of the represented domain (e.g. breast cancer). In contrast to groups (2) and (3), knowledge in a KB is not scattered over various representational spaces or layers, but integrated into one.

Knowledge levels, groups, or layers have of course been discussed previously in the AI literature. For instance, Newell introduced an agent-based distinction between the 'knowledge level' and the 'symbol level' in [23], and [1, 10, 11] analysed layers in formal ontology design. In more detail, Brachman, in 1979, introduced a classification of the primitives used in KR systems at the time [1], distinguishing the following four levels: (i) 'Implementational', (ii) 'Logical', (iii) 'Conceptual', and (iv) 'Linguistic'. Guarino [10, 11] added to these four layers yet another layer, namely the 'Epistemological Layer' for the primitives, situated between the 'Logical' and the 'Conceptual' layers. Our approach differs in that it mainly aims at distinguishing *human agents as individuals and groups* focused around particular epistemic interests, whilst analysing the corresponding impact on representation types. A more detailed analysis of the relationship to previous 'layering approaches' is left for future work.

[1] See http://www.ifomis.org/bfo/

[2] See http://www.onto-med.de/ontologies/gfo/

[3] See http://www.w3.org/TR/owl2-overview/

3 ONTOLOGY INTEROPERABILITY

We next discuss how the six epistemic groups impact on representation types, choice of formalisms, kinds of metadata, mappings, as well as reasoning. We begin by inspecting the notion of an ontology itself.

A plurality of ontologies and formalisms

An often cited definition of the term 'ontology' in computer science was given by Tom Gruber in 1992 [9] (here heavily abridged).

> A conceptualisation is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualisation, explicitly or implicitly.
>
> An **ontology** is an explicit specification of a conceptualisation. [...] For AI systems, "what exists" is that which can be represented. [...] In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory. [9, p. 908–909]

This definition, whilst being controversial, still nicely captures the main differences between the usage of the term 'ontology' in philosophy vs. computer science and artificial intelligence. Namely, consider the following snippets from this definition:

- 'simplified view of the world that we wish to represent for some purpose': an ontology as a technical artefact is not intended to cover the world in its entirety, but only chosen aspects of the world, on specific levels of abstraction, and for given purposes—largely independent of particular metaphysical positions such as realism and antirealism; here, group (4) will typically informally specify the relevant domain knowledge, whilst group (5) is in charge of establishing an agreement on how to formally codify this knowledge.
- 'committed to some conceptualisation': ontologies presuppose various decisions concerning ontological commitments. These originate partly in common sense knowledge (group (1)), precisifications given by members of group (2), and agreements as they are established in groups (3) and (4). Finally, the formal implementation of the ontological commitments is again left for groups

(5) and (6), merging collaborative interests of (1)–(6).

- "'what exists" is that which can be represented': ontological commitments are dependent on the expressive capabilities of selected representational formalisms. The choice of an adequate formal language can only be established as an interplay between logician (group (5)), computer scientist (group (6)), and the domain experts of (3) and (4).
- 'representational vocabulary' and 'human-readable text': there is a 'tension' between the logical vocabulary used, and the natural language concepts and terms it is meant to capture, and, in the case of e.g. breast cancer, various forms of scientific representations such as graphs, mathematical equations, images, 3D models etc. Reconciling this tension requires deep interaction between the various groups of domain experts and formal logicians and computer scientists.
- 'an ontology is the statement of a logical theory': on a technical level, an ontology is seen as equivalent to a logical theory, written in a certain formalism. Clearly, this task is for group (5), respecting the requirements of group (6).

Heterogeneity of formal languages is particularly important in the life sciences, where size of ontologies and needed expressivity vary dramatically. For example, whereas weak (i.e. sub-Boolean) DLs suffice for the NCI thesaurus (containing about 45.000 concepts) which is intended to become the reference terminology for cancer research [26], other medical ontologies such as GALEN[4] require the full expressivity of the OWL language (a decidable fragment of first-order logic), while foundational ontologies typically require at least full first-order logic (see [16]).

An example of a heterogeneous combination of formalisms is discussed in [13], where it is shown that in order to adequately represent the spatial structure of molecules as they are described in chemical ontologies such as ChEBI [2], ontology languages need to be combined with formalisms such as monadic second-order logic. We next investigate how such diversity and heterogeneity is reflected in and how it originates from the different group interests involved in the representation of breast cancer phenotypes.

A plurality of mapping and reasoning types

In biomedical ontologies, metadata in the form of tags, annotation, or more generally documentation, is of particular importance. Indeed, many biomedical

[4] See `http://www.opengalen.org/`

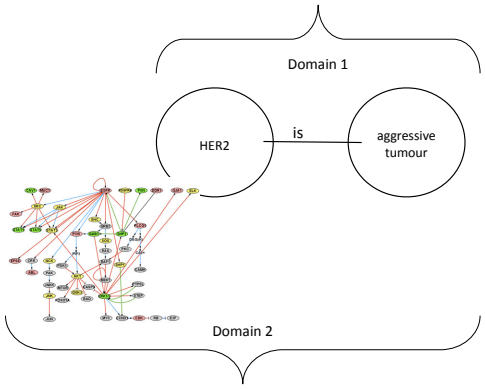

Fig. 1. Knowledge granularity.

ontologies have an extremely shallow logical structure, namely consist only of taxonomies, or even just of sets of concepts, however accompanied with a rich set of metadata. It is clear that the separation of the epistemic groups from Section 2 has a direct impact on the kinds of annotations and metadata that can be expected to be generated. For instance, the particular scientific communities (groups (2) and (3)) need not associate identical sets of concepts as related to a term in use. Had the 'Human Epidermal growth factor Receptor 2' (HER2, also known as ErbB2) been used as a tumour marker in the community of clinical oncologists, it would have been related to the diagnosis of an aggressive tumour with a poor clinical outcome and a low likelihood of a long term survival. On the other hand, among the group of molecular biologists HER2 would be associated with the specific protein-protein interactions that trigger the carcinogenic events.

As interests diverge among and within disciplines concerning ways of describing a problem, distinguishing similarities and difference makers will vary among knowledge domains. So, HER2 will not be the same difference maker for a clinician and for a biologist. The main difference that will be relevant for a clinician will be a difference in the patients survival associated with the expression of HER2 [27]. The biologist who focuses on the cellular signalling pathways might favour a differential expression of the ErbB2 gene while comparing the phenotypes of two types of cell lines [19]. Consequentially, justification of asserted similarities and generalisations will ask for a different kind of evidence in diverse domains. Clinical evidence will be acquired through survival analysis and clinical trials while biologists provide evidence through diverse experimental and explanatory methodologies [18]. Accordingly, the reasoning of

the groups (2)–(4) influence the related mappings and justifications implemented by the groups (5) and (6).

A relation between a term and its reference class gets its justification within domain knowledge as an adequate mapping relationship. The justification is expressed through the claims that support the mapping relations. Regarding the previous example, 'HER2' will be mapped onto a bad prognosis within clinical knowledge, and the mapping will be justified by the statistical data retrieved from the survival analyses (see Fig. 1, Domain 1). Likewise, biological knowledge provides an alternative mapping relation and a related justification to the mapping between 'HER2' and 'tumour aggressiveness', e.g. protein interaction pathways that result in cell proliferation and tumour aggressiveness (see Fig. 1, Domain 2). These diverse patterns of clinical and biomedical reasoning [3] can be perceived as domain specific. A detailed analysis of the mappings within and between knowledge domains asks for a multidisciplinary approach involving a community based process of knowledge production [5]. A group of experts with a common interest is collaborating in establishing standards that help them label and describe the domain of interest [20].

4 DISCUSSION AND FUTURE WORK

Concurrently with the systematisation of epistemic group levels, representation types and knowledge base types, we intend to use the introduced distinctions in order to characterise domain specific knowledge representations for breast cancer phenotypes. Specifically, we are interested in the problem of merging knowledge from different domains and in analysing the 'domain knowledge problems' of [14] further through inspecting a number of examples from molecular oncology and clinical practice. Here, we have demonstrated that such domain problems ask for a plurality of onto-logical formalisms.

We have sketched the intertwined processes involved in the integration of heterogeneous representations as they originate from different epistemic groups that are involved in complex domains such as breast cancer research. Concerning formal representations dealing with the heterogeneities of phenotypes, we propose to endorse a framework that allows to organise the various (domain) representations into an interlinked modular structure, respecting the plurality of formalisms, expressivities and aims, as they are found across diverse scientific communities. A further characterisation of the domain specific epistemic interests, including a deeper understanding of the characterised groups (1)–(6), would provide a more sustainable integration of knowledge about breast cancer, increasing interoperability of represented information and, therefore, applicability of acquired clinical and biological knowledge. A closer understanding of the domain needs would also further support decisions about which formalisms best suit a domain. [15, 22] lay the foundation for a distributed ontology language DOL, which will allow users to use their own preferred ontology formalism whilst becoming interoperable with other formalisms. At the heart of this approach is a graph of ontology languages and translations between them (see [17] for the theoretical development).[5] This graph enables users to:

- relate ontologies that are written in different formalisms with various kinds of mappings,
- re-use ontology modules even if they have been formulated in different formalisms, and
- re-use ontology tools like theorem provers and module extractors along translations.

Indeed, we believe that no attempt at an integration of knowledge can be epistemically sustainable unless it respects the plurality of formal languages and tools, methodologies and perspectives as they result from the heterogeneity of the domain interests.

ACKNOWLEDGEMENTS

Work on this paper was supported by the DFG-funded Transregional Collaborative Research Centre on Spatial Cognition (SFB/TR 8) and by the 'Fondazione Umberto Veronesi' (FUV). We are grateful for the very useful feedback of three anonymous reviewers.

REFERENCES

[1]R. J. Brachman. On the Epistemological Status of Semantic Networks. In N. V. Findler, editor, *Associative Networks: Representation and Use of Knowledge by Computers*. Academic Press, 1979.

[2]P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck. Chemical Entities of Biological Interest: an update. *Nucl. Acids Res.*, 38:D249–D254, 2010.

[3]A. D. Evans and V. Patel, editors. *Cognitive Science in Medicine: Biomedical Modeling*. MIT Press, Cambridge, MA, 1989.

[4]D. Faratian, R. G. Clyde, J. W. Crawford, and D. J. Harrison. Systems pathology: taking molecular

[5] DOL is currently under standardisation as Working Draft ISO/WD 17347 in ISO/TC 37/SC 3 'Systems to manage terminology, knowledge and content'.

pathology into a new dimension. *Nature reviews. Clinical oncology*, 6(8):455–464, 2009.

[5]J.-P. Gaudillière and H.-J. Rheinberger, editors. *From Molecular Genetics to Genomics : The Mapping Cultures of Twentieth-Century Genetics*. Routledge, London, 2004.

[6]R. Giere. An agent-based conception of models and scientific representation. *Synthese*, 172(2), 2010.

[7]A. Goldhirsch, W. C. Wood, A. S. Coates, R. D. Gelber, B. Thürlimann, and H. J. Senn. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer. *Annals of Oncology*, 2011.

[8]M. K. Gospodarowicz, B. O'Sullivan, and L. H. Sobin, editors. *Prognostic factors in cancer: International Union against Cancer*. Wiley-Liss, 2006.

[9]T. R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, 43(4-5):907–928, 1995.

[10]N. Guarino. The Ontological Level. In R. Casati, B. Smith, and G. White, editors, *Philosophy and the Cognitive Sciences*, pages 443–456. Hölder-Pichler-Tempsky, 1994. Proc. of the 16th Wittgenstein Symposium, Kirchberg, Austria, Vienna, August 1993.

[11]N. Guarino. The Ontological Level: Revisiting 30 Years of Knowledge Representation. In Alex Borgida, Vinay Chaudhri, Paolo Giorgini, and Eric Yu, editors, *Conceptual Modelling: Foundations and Applications. Essays in Honor of John Mylopoulos*, pages 52–67. Springer, 2009.

[12]I. Hacking. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, 1983.

[13]J. Hastings, O. Kutz, and T. Mossakowski. How to model the shapes of molecules? Combining topology and ontology using heterogeneous specifications. In *Proc. of the Deep Knowledge Representation Challenge Workshop (DKR-11), K-CAP-11, Banff, Alberta, Canada*, 2011.

[14]A. Hunter and R. Summerton. A knowledge-based approach to merging information. *Knowledge-Based Systems*, 19(8):647–674, 2006.

[15]O. Kutz, T. Mossakowski, C. Galinski, and C. Lange. Towards a Standard for Heterogeneous Ontology Integration and Interoperability. In *International Conference on Terminology, Languages and Content Resources (LaRC-11)*, Seoul, South Korea, 2011.

[16]O. Kutz, T. Mossakowski, J. Hastings, A. Garcia Castro, and A. Sojic. Hyperontology for the Biomedical Ontologist: A Sketch and Some Examples. In *Workshop on Working with Multiple Biomedical Ontologies (WoMBO at ICBO 2011)*, Buffalo, NY, USA, August 2011.

[17]O. Kutz, T. Mossakowski, and D. Lücke. Carnap, Goguen, and the Hyperontologies: Logical Pluralism and Heterogeneous Structuring in Ontology Design. *Logica Universalis*, 4(2):255–333, 2010. Special Issue on 'Is Logic Universal?'.

[18]A. La Caze. The role of basic science in evidence-based medicine. *Biology and Philosophy*, 26(1):81–98, 2011.

[19]M. Lacroix and G. Leclercq. Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast cancer research and treatment*, 83(3):249–289, 2004.

[20]S. Leonelli. Bio-ontologies as Tools for Integration in Biology. *Biological Theory*, 3(1):7–11, 2008.

[21]C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. WonderWeb Deliverable D18: Ontology Library. Technical report, ISTC-CNR, 2003.

[22]T. Mossakowski and O. Kutz. The Onto-Logical Translation Graph. In *Modular Ontologies—Proc. of the Fifth International Workshop (WoMO 2011)*, volume 230 of *Frontiers in Artificial Intelligence and Applications*, pages 94–109. IOS Press, 2011.

[23]A. Newell. The Knowledge Level. *Artifical Intelligence*, 18(1):87–127, 1982.

[24]J. D. Novak and A. J. Cañas. The Theory Underlying Concept Maps and How to Construct Them. Technical report, Florida Institute for Human and Machine Cognition, 2008.

[25]C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A.-L. Borresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[26]N. Sioutos, S. de Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, and L. W. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.

[27]D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235(4785):177, 1987.