Enhancing biological information systems with granularity

C. Maria Keet
Faculty of Computer Science, Free University of Bozen-Bolzano, Italy. keet@inf.unibz.it

1. Introduction

The data explosion in (applied) biology and its associated data management problems of Internet-available resources is well-known (e.g. [3]), but how to solve it remains a big question. The recently established Health Care and Life Sciences SIG of the W3C [24] aims to contribute to the creation of a Life Sciences Semantic Web (LSSW) by applying existing Semantic Web Technologies, based on the assumption that what is good for enterprise modelling will do for biology as well. To realise the LSSW, one needs a least to link the many data repositories to be able to find the right data or information at the right level of detail quickly and at once. For biology, this requires *vertical* integration of domain knowledge from fine-grained genotype up to environmental living conditions to relate, among others, epidemiology, healthcare statistics, physiology, and genetics to discover e.g. a gene or gene complex that causes a disease or contributes to an aspect of longevity. Semantic Web agents can negotiate about the meaning of the data sources through services such as WSDL, provided the semantic descriptions are available. This generation and use of annotation of biological resources, a prerequisite for realising the LSSW, is a crucial open problem. Formal ontology and ontologies represented in OWL or RDF are useful generic methodologies to deploy, but they do neither address the different levels of biological granularity *per se* nor take into account that users with different foci will have to use them. Granularity is already a pressing problem to address in biological data management even in absence of a LSSW, and it worsens as biologists keep rapidly producing more data without better data- and information management. Semantic Web Technologies provide methodologies that can be used for developing a solution and integrate granularity-based data and information management to exploit existing data better and bring more structure to the 'biological data anarchy' on the Web. However, a) biological granularity has not been subjected to a formal analysis yet, b) the nature of biological data & information is more diverse, hence challenging, than e.g. operational data of enterprises, and c) how granularity can be applied as an additional ontology-like formal structure and be used by Semantic Web services, querying, reasoning, and information visualisation is underspecified by biologists, ontologists, and engineers alike.

2. Current approaches to granularity

To the best of my knowledge, there are no existing approaches combining granularity, (biological or otherwise) and the Semantic Web, therefore I focus here on granularity.

Tange *et al* [16] constructed informal shallow granular perspectives and levels for text mining and categorization of scientific literature about medical practice, which is the only topically-related software application. Two other software applications that use granularity are student-tailored study feedback [11] and computer games [22]. Interestingly, [16] [22] achieved better performance with multiple shallow granularity hierarchies than with fewer perspectives that had more detailed levels. Other informally established granular levels include the omics spaces (levels) [17] [23], which do not contain one type of relation between them (they are related through time displacement and causality, roughly). On the border of biomedicine and formal implementations is Kumar *et al*'s simple *gran* function [9], where

GR is the ordered set of levels of granularity applicable to a domain and *U* denotes the set of biological universals such that $gran : u \rightarrow gran(u)$, for $u \in U$ and $gran(u) \in GR$ and returns the level of granularity where the entity resides. The basic idea is useful, but they have not tested it and it has several shortcomings: it assumes that the domain has been partitioned already (but how are the entities assigned to a level?) and it requires patchwork in the logic, design and implementation. For instance, *gran* returns only one level at a time, which cannot be scaled up to use it in a subject domain with more than one perspective without redefining the function to return more than one level of a given entity ([5] provides an unsatisfactory example) or requires an *ad hoc* identifier management system. Another limitation, which is equally problematic with e.g. [17] [15], is the bottom-up development of granular levels limited to human structural anatomy [9] [14], which is not reusable in a larger subject domain such as infectious diseases [5]. For computational implementations, however, an underlying domain-independent logically consistent theory of granularity is an imperative to meet requirements such as reusability, flexibility, and interoperability. Aside from Hobbs' formalised sketch of granularity [2], Bittner and Smith [1] have developed a more comprehensive "theory of granular partitions", which is limited to mereology, does not address the types of aggregation commonly used with data mining and conceptual modelling, nor linking of perspectives (partitions), and does not allow for the kind of granularity and abstraction commonly used in biology [4] or linguistics [10].

Other types of implementations exist in different research disciplines, such as data mining and clustering techniques, which are grouped recently under the term *Granular Computing*. It combines efforts from philosophy, AI, machine learning, database theory and data mining, (applied) mathematics with fuzzy logic and rough sets, among others [19], for example [18] [13] [20] [21] [12], where its characteristic computational problem solving emphasises *quantitative* aspects of granularity over philosophical aspects that stress the *qualitative* component. These are largely separate research communities with little cross-fertilisation, but could work in synergy to enhance both approaches toward granularity.

Abstraction is related to granularity, which is the *process* to go from a detailed to simplified representation – i.e. going from fine to coarse-grained levels of granularity. Without pre-defined levels, application support for abstraction is limited to manual procedures, syntax, and heuristics [4].

3. Proposed solution

3.1 The vision

The aim of this research is to develop a formal, domain- and implementation-independent theory of granularity that can, and will, be used for computational reasoning in different subject domains. This theory is orthogonally positioned, added to, the data sources (data- and knowledge bases and ontologies) to enhance data and information management and inferencing across levels of granularity. The theory and prototype implementation will be sufficiently comprehensive to be useful in the subject domain of biology. The overview of the basic architecture is depicted in *Fig 1*. The source data on the left-hand side provides the content that is loaded into the domain granularity framework, or the data source receives an additional layer of logic to enhance querying and inferencing. The meta-granularity is on the right-hand side, where it is specified what the granularity components are, how adjacent levels relate to each other, and how the components – level, perspective, domain – are related. The definitions and constraints at this meta-level restrict the specification of each domain granularity framework (in the centre of the figure) to ensure consistency of the domain perspectives and levels with the theory. These three major components together will result in a robust characterisation and implementation of granularity.

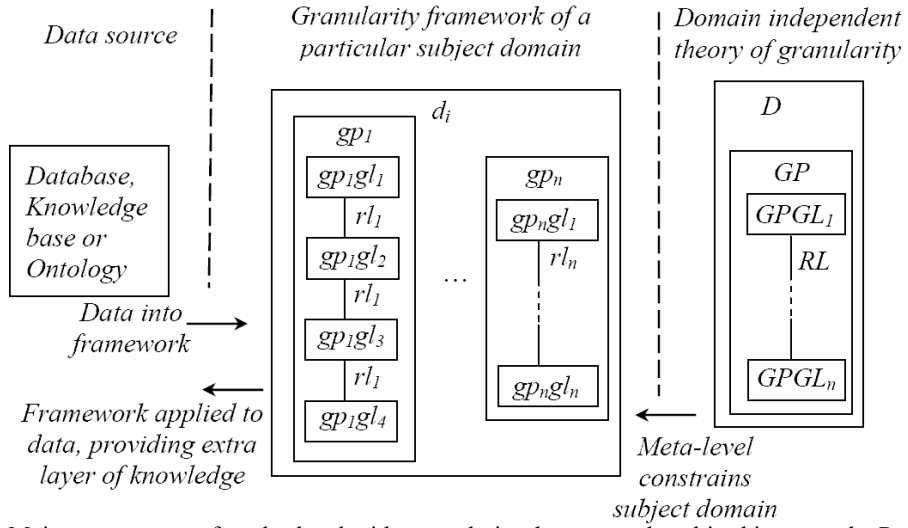

Figure 1. Main components of, and related with, granularity that are analysed in this research. *D*: domain, *GP*: granular perspective, *GPGL*: granular level of a perspective, *RL*: relation between two adjacent levels, d_i: subject domain, gp_i: perspective in the subject domain, gp_1gl_1... gp_ngl_n: levels defined for each perspective, rl_i: relation between two levels in the domain granularity framework.

3.2 Achievements and planned research

The main theoretical contributions I have made to date to realise this vision, are the disambiguation of types of granularity into a taxonomy and data structure internal to granular levels [6], ontological considerations, a FOL characterisation of the domain independent theory of granularity and its data manipulation operators, including abstraction [4]. On the experimental side, informal [5] and formal perspectives, levels, and data manipulation operators for the infections diseases domain were defined and tested using DL [8]. Granular information retrieval was tested with the Foundational Model of Anatomy and the Gene Ontology and its limitations concerning representation and implementation is discussed [7].

Planned research involves: 1) refining the FOL characterisation of the theory and providing a more comprehensive justification of the modelling considerations and decisions (indistinguishability, and formal aggregation/part-of, 2) improve feasibility for data manipulations (e.g. recursive queries), ontology linking, 3) test cases with infectious diseases and nuclear hormone receptors, and 4) practicability with DL, OWL, RDF and SPARQL.

4. Comparison proposed solution to existing approaches

Summarizing the three groups of shortcomings in existing approaches, described in §2, to address the problems briefly outlined in §1, which the solution proposed in §3 shall solve:

A. The informal approaches are not usable for computation and reasoning, are (philosophically) inconsistent, and underspecified.

- ***Solution***: Development of a domain-independent theory of granularity, comprising disambiguation between scale- and non-scale-dependent types of granularity and ontologically motivated modelling decisions necessary for the development of a generic, ontologically sound formal foundation of granularity, which enables automated reasoning and precise specification of domain granularity.

B. The formal approaches are more or less compatible *partial* solutions to solving what granularity is and how to use it, are not context-aware, and do neither have formalized nor tested what one can do with it, therefore they are only of limited use.

- ***Solution***: Foundational semantics, formal representations, and operators are related through a unifying theory that maps to a model-theoretic semantics usable to specify and

constrain domain granularity frameworks, which is necessary for computation and interoperability. The data manipulators enable querying and reasoning to manage the databases, ontologies-stored-in-databases, knowledge bases, and (OWL-)ontologies to which granularity is applied, providing a novel method to analyse information as opposed to only describe it.

C. The engineering solutions are not reusable in the current format beyond the software application each one is designed for, with the familiar problems in interoperability.

- ***Solution***: The domain and implementation independent theory of granularity ensures its genericity and widest possible applicability such that multiple divergent uses have a shared common well-founded framework. For instance, the research will be useful in (biological) database management, cross-granularity querying and reasoning, ontology and pathway management with 'browsing in context', fact finding, and provide support for other computational implementations like information retrieval from large corpora.

5. Benefits of the research

The combination of maximum expressivity and computability takes advantage of both foundational ontology aspects and engineering usefulness for implementations. The domain- and implementation-independence and formal rigour will be more comprehensive than abstraction and granularity separately achieve and can be applicable to a wider variety of software applications, thereby increasing usability and reusability. There is limited, *ad hoc*, 'context aware' ontology browsing and agent mediation that is difficult to scale-up, but will be much easier with the proposed theory. It can make the huge amounts of data and large ontologies in the LSSW manageable and understandable because the user will be able to 'zoom in' to the desired section & level(s), hiding irrelevant information, yet at the back-end this is still linked and usable for inferencing and for other users with different foci. The additional logic layer integrated with Semantic Web technologies enables data manipulation through recombination of the information to retrieve more information from the same data source, where discovering gaps and errors in the data source can be easily found and corrected or provide an impetus for wet-lab research.

6. Research methodology

Implicitly, the aims in §3.1 contain the hypothesis that there exist a 'granular view' on reality – or, more strongly: that reality is granular – that can be represented in one encompassing theory. Consequently, it may be falsified if a) the theory of granularity requires exceptions for each subject domain – suggesting that there is no underlying framework after all, or the theory is insufficient – and b) by the inadequacy of applicability to any subject domain – indicating that the theory is not properly grounded. To test the hypothesis, I do not take a pure top-down or bottom-up approach, but follow one that is mainly top-down with several bottom-up examples taken primarily from the biology subject domain in order to provide intermediate feedback during definition of the theory and applicability of the theory (see also achievements described in §3.2).

First, a preliminary investigation comprised a literature and application search, including a possible solution extending contextual reasoning, and experimenting with domains like infectious diseases and some examples in ecology. This aided understanding of perspectives, their criteria, levels, and formulation of research questions. Second, the coarse-grained formulation of research questions (sub-questions and tasks are omitted due to space limitations) that will be answered through iterations between developing a formal theory and experimentation with biology case studies:

1. Can a domain-independent theory of granularity be defined and formalised?
2. How can this theory of granularity constrain domain specific granularity?
3. How to load/apply domain data in/to the domain specific granularity framework?
4. What reasoning tasks can benefit from granularity? Where and how will this affect identified types of tasks and development of a model theory?

References

[1] Bittner, T., Smith, B. A Theory of Granular Partitions. In*: Foundations of Geographic Information Science*, Duckham, M., Goodchild, M.F., Worboys, M.F. (eds.), London: Taylor & Francis Books, 2003, pp117-151.

[2] Hobbs, J.R. Granularity. International Joint Conference on Artificial Intelligence, 1985. pp432-435.

[3] Keet, C.M. Current characteristics and historical perspective of Computer Science and IT with/for biology. In: *CSBio reader: extended abstracts of the 'CS&IT with/for biology' Seminar Series*, Free University of Bozen-Bolzano. Keet, C.M., Franconi E. (eds.). 2005. pp1-8.

[4] Keet, C.M. Using abstractions to facilitate management of large ORM models and ontologies. Int'l Workshop on Object-Role Modeling (ORM'05). In: *OTM Workshops 2005*. Halpin, T., Meersman, R. (eds.), LNCS 3762. Berlin: Springer-Verlag. pp603-612.

[5] Keet, C.M., Kumar, A. Applying partitions to infectious diseases. *XIX Int'l Congress of the European Federation for Medical Informatics (MIE2005)*, 28-31 August 2005, Geneva, Switzerland. 2005.

[6] Keet, C.M. A taxonomy of types of granularity. *IEEE Conference in Granular Computing (GrC2006)*, 10-12 May 2006, Atlanta, USA.

[7] Keet, C.M. Using bio-ontologies in RDBMSs for querying granular information. *Bioinformatics, Ontologies and databases (SBIOLBD 2006)*, 13-2-2006, EPFL, Lausanne, Switzerland. (extended abstract).

[8] Keet, C.M. A formal approach for using granularity in the subject domain of infectious diseases. Technical Report KRDB06-2, Faculty of Computer Science, Free University of Bozen-Bolzano, 2006

[9] Kumar, A., Smith, B., Novotny, D.D. Biomedical Informatics and Granularity. *Comp. & Funct. Gen*., 2005, 5(6-7): 501-508.

[10] Mani, I. A theory of granularity and its application to problems of polysemy and underspecification of meaning. In: *Principles of Knowledge Representation and Reasoning: Proc. of the 6th Int'l Conference (KR98)*, A.G. Cohn, L.K. Schubert, S.C. Shapiro (eds.). San Mateo: Morgan Kaufmann. 1998. pp245-255.

[11] McCalla, G.I., Greer, J.E., Barrie, B., Pospisil, P. Granularity Hierarchies. *Computers and Mathematics with Applications: Special Issue on Semantic Networks*, 1992, 23: 363-376.

[12] Peters, J.F., Skowron, A., Ramanna, S., Synak, P. Rough sets and information granulation. In: T.B. Bilgic, D. Baets, and O. Kaynak (eds*.), Proc. of 10th Int'l Fuzzy Systems Association World Congress*, LNAI 2715, Springer-Verlag, 2002. pp370-377.

[13] Reformat, M., Pedrycz, W., Pizzi, N. Building a software experience factory using granular- based models. *Fuzzy Sets and Systems*, 2004, 145:111-139.

[14] Rosse, C. and Mejino, J.L.V.: A reference ontology for biomedical informatics: the foundational model of anatomy. *J. Biomed. Inf.,* 2003, 36: 478-500.

[15] Stell, J.G. Qualitative Extents for Spatio-Temporal Granularity. *Spatial Cognition and Computation*, 2003 3(2-3): 119-136.

[16] Tange, H.J., Schouten, H.C, Kester, A.D.M., Hasman, A. The Granularity of Medical Narratives and Its Effect on the Speed and Completeness of Information Retrieval. *J. Am. Med. Inf. As*., 1998, 5(6): 571-582.

[17] Toyoda, T. and Wada, A. Omic space: coordinate-based integration and analysis of genomic phenomic interactions. *Bioinformatics*, 2004, 20(11): 1759-65.

[18] Yao, Y.Y. A partition model of granular computing. *LNCS Transactions on Rough Sets*, 2004, 1:232-253.

[19] Yao, Y.Y. Perspectives of Granular Computing. *IEEE Conference on Granular Computing* 2005,

[20] Zadeh, L.A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 1997, 90(2):111-127.

[21] Zhang, J., Silvescu, A., Honavar, V. *Ontology-Driven Induction of Decision Trees at Multiple Levels of Abstraction*. Technical Report ISU-CS-TR 02-13, Computer Science, Iowa State University. 2002.

[22] Zukerman, I., Albrecht, D., Nicholson, A., Doktor, K. Trading off granularity against complexity in predictive models for complex domains*. In: Proc. of the 6th Int'l Pacific Rim Conference on Artificial Intelligence (PRCAI 2000)*.

[23] Infobiogen. http://www.infobiogen.fr/deambulum/tab.php?page=vue&lg=en.

[24] W3C Semantic Web Health Care and Life Sciences Interest Group. http://www.w3.org/2001/sw/hcls/.