Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220720

The Sortal Concept in the Context of Biomedical Record Linkage

Marko MILETIC[a] and Murat SARIYAR[a,1]
[a] Bern University of Appl. Sciences, Switzerland

Abstract. Biomedical Record Linkage is especially designed for linking data of patients in different data repositories. An important question in this context is whether singling-out is sufficient for identifying a patient, and if not, what is in general required for identification. To provide hints for an answer, we will extend previous works on the concept of identity and extend the sortal concept, stemming from analytical philosophy and upper-level ontologies. A sortal is a concept that is associated with an identity criterion. For example, the concept "set" has the identity criterion "having the same members". Based on a description of a record linkage setting, we operationalize the sortal concept by providing a distinction between the digital representation of a person (d-sortal) and the person in flesh (b-sortal).

Keywords. Identity, record linkage, reference reconciliation, sortal, strawson

1. Introduction

Biomedical record linkage (RL) is especially designed for linking data of patients in different data repositories [1,2]. A significant question in this context is whether singling-out is sufficient for identifying a patient, and if not, what is generally required for identification. In a previous work, we set the ground for an ongoing investigation into such questions and started with a differentiation between different sorts of identity (numerical, qualitative, and relational identity [3]). Here, we extend this work by introducing the sortal concept into the RL domain. These two have the goal to solidify the ground for assessing as to when identification is really achieved through RL. As they translate philosophical concepts into an application field, abstract considerations cannot be avoided and will be further concretized in the coming empirical work.

Originating in analytical philosophy by reference to its first usage by Locke, the sortal concept is primarily used in philosophy, logics and knowledge representation [4]. Using a simplified definition, a sortal is a concept that is associated with an identity criterion. For example, the concept "set" has the identity criterion "having the same members". More formally and according to the version of Peter Strawson, a concept F is a sortal iff the following formula holds for a suitable equivalence relation R [5]:

$$\forall x \forall y \, ((F(x) \wedge F(y)) \rightarrow R(x,y) \leftrightarrow x=y) \quad (1)$$

[1] Corresponding Author, Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

The equivalence relation R in (1) indicates how the concept (usually a property) translates into the identity of F-kind entities. Two sortals with a broad R are materials and events: for the former one, a possible identity criterion is to occupy the same place at any time, and for the latter, an identity criterion could be to have the same causes and effects. For record linkage in health contexts, three questions arise: (i) what should F be (e.g., "person")?; (ii) what should R be (e.g, "having the same DNA sequence")?; and (iii) what is the benefit for the practice of using the sortal concept? Our innovation is to provide an orientation sortals that is not captured in existing ontologies, and that will inform the practice and further methodological development in biomedical RL.

For example, the Unified Foundational Ontology (UFO [6]), extending and unifying the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE [7]) and the General Formal Ontology (GFO[8]), uses the sortal concept to distinguish those substantial endurant types with identity criteria from those that do not have one. However, from the perspective of solving practical issues of identification, UFO is not sufficient due to two important gaps. First, identification is taken for granted and does not represent a subject to be dealt with. Second, crucial investigations on identity in the philosophical literature are not considered. For example, philosophical views on relative identity (sortal-relativity) or rigid designators conducted by Deutsch, Geach, Hirsch, Noonan, Strawson, Kripke, Lowe, Wiggins, and many more, are not visible, leading to the necessity of implicit decisions. UFO, for instance, does not allow an individual to instantiate more than one kind, which could be read in favor and against relative identity (depending on the concretization of the concept "individual").

In the following, we will first introduce general concepts to operationalize the sortal concept and provide a description of a record linkage setting, after which we address the questions (i)-(iii) by referring to this setting. We will conclude with some implications for the practice and further research.

2. Methods

To tackle the question as to how the results of record linkage must be enriched to grasp the concrete person behind data records, sortal-relativity is assumed (following Deutsch [9] and Lowe [10]), as the identity of a person can mean different things, for example, the body, the id card, or personality. The identity criterion of these different sortals must be detailed out to be meaningful for identification. For example, personality does not have an identity criterion, which makes it impossible to decide whether in borderline discussions such as brain damages, the personality remains the same or not. On the other hand, reference to the body can be associated with an identity criterion such as "having the same DNA fingerprint", but might not be sufficient under certain circumstances, for instance, for associating a blood sample to a patient. Hence, the decision which sortal to use has significant impacts on the practice of record linkage.

In our theoretical record linkage setting, we assume that data of two registries, a cancer registry and a register of residents, are to be linked to track the survival of patients. When performing probabilistic record linkage with two thresholds, probable matches are allowed, which must be resolved manually. The data items used for linkage are (firstname, lastname, place, dbirth, mbirth, ybirth, sex), and in the manual inspection further items can be required from the residents' office: full address and last contact. Let's assume that a probable match occurs because of misspellings in the first three data items. Now, having this setting, how to guide the manual inspection and classification

by referring to the sortal concept? We will discuss how the different perspectives on a person can be operationalized with the help of sortals, when the cancer registry contacts the register of residents for resolving possible matches.

3. Results

For tracking the survival, a sortal with the identity criterion "having the same values of a set of variables" seems sufficient, as no contact to the underlying person is necessary. Now, the important question is: to what kind of sortal (F) does this correspond to? Two possible answers are "digital representation of a person" (d-sortal) and "digital representation of properties of a person". From a layman point of view, the implications seem identical. However, with the former proposal, it is possible to exploit external relations of a person. For example, the person who issued an ID card together with the place and the date, is not a property of a person, but can be part of her representation.

Biomedical RL just relies on the d-sortal for deciding identity, and it might seem obvious, that it is sufficient for manually resolving possible matches as well. This is true, for instance, when there are misspelling that can be resolved with or without the help of the additional attributes, e.g., as for the following data pair: ('Urs', 'Schmidt', 'Bern', '18', '11', '1990', 'm') and ('Urz', 'Schmitt', 'Berne', '18', '11', '1990', 'm').

Even in such cases, there is no guarantee to manually assign the correct matching status, as the d-sortal does not guarantee an arbitrarily increaseable granularity of the properties used. What is often meant by manual review is the inclusion of plausibility considerations that increases the granularity in order to better represent the real individual, which is represented by a sortal (let us call it b-sortal) that has an identity criterion such as "having the same iris", "having the same fingerprint" or "having the same flesh-and-bone-body". An epistemological problem of the b-sortal is the difference between discrete contents of thought and the continuous aspects or changes of the real world, which can only be addressed by physical presence of the individual and a rigid association of this presence with a signature of that very individual or a recorded validation, signed by another individual who can be trusted. As this is in most cases not feasible for biomedical RL, another approach must be pursued.

To grasp the relation between the two sortals and develop a solution for the epistemological problem, an analogy with the famous lump-statue case can help: even though a statue of Platon might be made by the same lump of bronze as a statue of Socrates, they are different according to the identity criterion of statues, which refers to the form and the material, not only to the latter one. The hard part of statue identification is the overlapping material, not the form. A d-sortal entity is like the form of a statue that is built upon the real person (lump). In contrast to the lump-statue case, the overlap between the two sortals is only virtual (both must be represented) and not material, which has the disadvantage, that the real entity cannot help in identification problems on the d-sortal level. On the other hand, one important advantage is the possibility to enlarge the overlap by means of further properties and especially relations. The latter ones increase the granularity of the d-sortal due to the sheer number of possible combinations, e.g., regarding the person who issued an ID card, the place and the date. Hence, relational identity is a promising solution that references more aspects of the b-sortal entity.

As a first step towards an ontology for identification and identity in the context of biomedical RL, these thoughts have practical implications in their own. First, as RL only deals with d-sortals, one cannot resolve possible matches by reference to the real

underlying person. Many circular definitions for identity and identification lack this insight. To use a real underlying person for resolving synonyms and homonyms would require to identify the person in the first place. If the variables used for identification do not have enough granularity, there is no way to compensate for it, and additional variables only help if they are available at the cancer registry as well as at the register of residents (e.g., the last contact information only helps if the cancer registry can perform plausibility checks due to survival data available). Second, the equivalence relation R in the identity criterion of the d-sortal should cover relational aspects to increase the granularity in order to better represent the real individual. Such relations should also cover synchronization of the two sortals, allowing to detect possible discrepancies between the representation and the real individual from time to time, which is a reason for the necessity to renew ID cards at certain time intervals. Third, the question whether singling-out is the same as identification cannot be answered without clarifying which sortals is meant. In practice, sortals are often mixed.

4. Discussion

Our results are first steps for developing an ontology for identification in biomedical RL. The authors have a high level of immersion in real-world concerns and introduced the sortal concept because it allows to clarify misconceptions and to develop more suitable RL methods based on a forthcoming ontology. Besides the basic sortal concept and relational identity, further concepts from the philosophical domain will be included in the ontology, e.g., phased sortals, perdurants, and rigidity. All these efforts will also rely on two upper-level ontologies central for us: UFO [11] and BFO [12].

Regarding the term d-sortal, it should be noted that it can be further concretized by including a list of attributes and relations for the identity criterion. Thereby, one would account for achieving anonymized data, e.g., by perturbing values in this list in such a manner that no data record unique. However, in a dynamic environment, it seems impossible to fix the list of attributes and to decide when data do not represent individuals anymore. This issue and its relations to the d-sortal suggest that anonymization can also heavily benefit from a sortal-based ontology for identification.

References

[1] Schouten LJ, Schlangen JT, de Rijke JM, Verbeek AL. Evaluation of the effect of breast cancer screening by record linkage with the cancer registry, the Netherlands. Journal of Medical Screening. 1998 Mar 1;5(1):37-41.
[2] Roos LL, Mustard CA, Nicol JP, McLerran DF, Malenka DJ, Young TK, Cohen MM. Registries and administrative data: organization and accuracy. Medical care. 1993 Mar 1:201-12.
[3] Sariyar M, Holm J. On the Concepts of Identity and Similarity in the Context of Biomedical Record Linkage. Stud Health Technol Inform 2021;281:472–476.
[4] Lowe EJ. Sortals and the Individuation of Objects. Mind Lang 2007;22:514–533.
[5] Strawson PF. Entity and Identity: And Other Essays. Oxford: Clarendon Press 2000.
[6] Guizzardi G, Wagner G. Using the Unified Foundational Ontology (UFO) as a Foundation for General Conceptual Modeling Languages. In: Poli R, Healy M, Kameas A (eds) Theory and Applications of Ontology: Computer Applications. Dordrecht: Springer Netherlands 2010;175–196.
[7] Borgo S, Ferrario R, Gangemi A, Guarino N, Masolo C, Porello D, Sanfilippo EM, Vieu L. DOLCE: A descriptive ontology for linguistic and cognitive engineering. Applied Ontology. 2022(Preprint):1-25.

[8] Herre H. General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling. In: Poli R, Healy M, Kameas A (eds) Theory and Applications of Ontology: Computer Applications. Dordrecht: Springer Netherlands 2010;297–345.
[9] Deutsch H. Identity and General Similarity. Philos Perspect 1998;12:177–199.
[10] Lowe EJ. Sortals and the Individuation of Objects. Mind Lang 2007;22:514–533.
[11] Amaral G, Baião F, Guizzardi G. Foundational ontologies, ontology-driven conceptual modeling, and their multiple benefits to data mining. WIREs Data Min Knowl Discov 2021;11:e1408.
[12] Arp R, Smith B, Spear AD. Building Ontologies with Basic Formal Ontology. MIT Press 2015.