ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Representing the Reality Underlying Demographic Data

William R. Hogan, Swetha Garimalla, Shariq A. Tariq

Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

Abstract. Demographic data about patients, research subjects, students and trainees, physicians and other healthcare providers, and so on is extremely important for nearly every biomedical application that manages information about people. The importance of demographics extends beyond biomedical informatics and touches fundamentally nearly every software application that manages information about people. However, we show that the treatment of demographic data in current information systems is ad hoc, and current standards are insufficient to support accurate capture and exchange of demographic data. We propose a solution based on realist ontology and implemented in the Demographics Application Ontology, which draws terms from reference ontologies such as the Phenotypic Quality Ontology and the Ontology of Medically Related Social Entities. Furthermore we have a created a web site that demonstrates the approach.

Keywords: Realist ontology, demographics, referent tracking

1 Introduction

Demographic data at their essence are data about people. There is no consensus set of "characteristics" that comprises demographics, but typically demographics include birth date, gender, sex, marital status, race, and ethnicity. Demographics are ubiquitous in software applications in healthcare and beyond, with obvious importance, for example, to electronic health records (EHRs), to the United States Census, and to the finance and retail industries. Demographic data are used to identify people, to make statistical comparisons of population groups, and to link records from multiple databases about one person.

At the University of Arkansas for Medical Sciences, we are studying Referent Tracking (RT) [1] as applied to EHRs. Clearly, we need to include demographic data in RT Systems (RTSs). As we shall discuss, we found current approaches to (and standards for) demographics inadequate. Therefore, we analyzed the reality on the side of the person that demographic data are about. On the basis of our analysis we propose a realist approach to demographics.

2 Preliminaries

Because there is no standard set of demographics, we had to choose one to start. Here, we discuss our rationale for the set we chose. We also set the stage for a review of current approaches to demographics by clarifying sex vs. gender.

2.1 Demographics Addressed in this Work

We started with the set of demographics required by the "meaningful use" (MU) regulation for EHRs in the United States (U.S.) [2], because we are studying RT applied to EHRs. For "eligible providers", this set includes preferred language, gender, race, ethnicity, and date of birth. For "eligible hospitals", it additionally includes date and preliminary cause of death (in the event of mortality in the hospital).

We then modified this set for pragmatic reasons. We *included* sex and marital status. We included the former because current approaches confuse it with gender (as we illustrate next). We included the latter because, in the U.S. at least, marriage confers on one's spouse broad healthcare visitation and (when one is incapacitated) decision-making rights. Providers therefore require marital status to know whose instructions to follow when the patient is incapacitated. We *excluded* preferred language, race/ethnicity, and preliminary cause of death. The first requires ontological theories of preferences and human language that are beyond our scope here. The second requires a treatment that is ongoing work, and for which we had insufficient space here to do justice. The third we assume can be handled by ongoing work in the ontology of disease and injury (e.g., Scheuermann et al.[3], Hogan [4], Cowell and Smith [5], and Goldfain et al. [6]).

2.2 Sex vs. Gender

The World Health Organization (WHO) explains the distinction between sex and gender thusly:[1]

"Sex" refers to the biological and physiological characteristics that define men and women. "Gender" refers to the socially constructed roles, behaviours, activities, and attributes that a given society considers appropriate for men and women.

Sex is biological; gender is psychosocial. In practice, the correlation between them is high, but there are transgendered and transsexual people who we will need to represent correctly in EHRs for optimal patient care and research.[2] Furthermore, with respect to sex, we can distinguish phenotypic and chromosomal (karyotypic) sex. Although they too correlate highly, there are individuals, for example, with an XY karyotype who are phenotypically female [7].

Driven largely by administrative requirements (i.e., billing), almost all structured healthcare data at present include only gender.[3] However, as we move towards a quality-driven healthcare system, quality of EHR data will become increasingly important, and thus also will, we believe, correct distinctions among gender, phenotypic sex, and karyotypic sex.

3 Limitations of Current Approaches

Here, we review usual approaches to demographics and standards for them. We divide these approaches into three major groups: "Person table", terminology standards, and semantic web.

3.1 The "Person Table" (or Class)

The typical approach to demographic data, especially in software that uses a relational database for persistence, is to have a "Person" table with fields for birth date, marital status, etc. Formal information models do little more than more than formalize the Person table as a class, and the fields as "attributes." Thus, the HL7 Reference Information Model (RIM) has a Person class, which "specializes" the LivingSubject class. The latter has an "attribute" called 'administrativeGenderCode',[4] and the former has "attributes" called "maritalStatusCode", "raceCode", etc.

The limitation of this approach is that it treats gender, date of birth, and marital status in exactly the same way. That is, to the computer the only difference between gender and date of birth is the field name and datatype: else they are the same type of entity ("attribute") related to the person in exactly the same way ("attribute of"). The true relationship between a human being and his or her gender, marital status, etc. is implicit and obscured.

3.2 Terminology Standards

The information model approach specifies that some attributes take a coded value. For example, the HL7 RIM has the codes M, F, UN as allowed values for "administrativeGenderCode", where UN stands for "undifferentiated". HL7 describes UN further by saying: *The gender of a person could not be uniquely defined as male or female, such as hermaphrodite.* And here we see that HL7 is confused, as 'hermaphrodite' refers to anatomical considerations and therefore sex, not gender. The RIM has no attribute for sex.

Terminology standards also confuse sex and gender. SNOMED CT (SNCT) places *Male* and *Female* as children of *Finding of biological sex*, but has no representation of male and female gender. It is thus either incomplete or assumes the sex codes are sufficient for gender too.

[1] http://www.who.int/gender/whatisgender/en/index.html

[2] There was a discussion on the listserv of the Association of Medical Directors of Information Systems (AMDIS) where one physician commented that he discussed with his patient how, in the future, he would have to "flip" the patient's gender in the registration system to ensure correct payment for prostate cancer screening (male-to-female transsexual surgery leaves the prostate in place and intact). For more on the AMDIS thread, see author WRH's discussion at http://hl7-watch.blogspot.com /2010/12/demographics-hl7-vs-reality-part-2.html

[3] MU also requires gender only, and not sex.

[4] Given that the primary use case for this attribute is bed assignment, we wonder why a "bed assignment instruction" code is not more apropos.

Furthermore, it also has *2547121016 Gender determination by chromosome analysis*. Clearly, we cannot ascertain one's socially constructed roles from chromosomal analysis; karyotype is what is meant here. The UMLS also confuses sex and gender, mapping SNCT codes for sex to UMLS concept unique identifiers for gender (e.g., *Male* to *C0024554 Male gender*).

LOINC has two codes (11882-8, 11883-6) for the sex of a fetus as observed on ultrasound, but labels the measured attribute as 'gender'. It would be hard to ascertain, even by ultrasound, the socially constructed roles of a fetus. Both codes have a "related name" of 'fetal sex'.

The NCIT has *Male gender*, but not as a subtype of *Gender*, oddly. Instead, it is a subtype of *Male*, itself a subtype of *"General qualifier"*. The siblings of *Male gender* include *Male phenotype* and *XX male*. The NCIT therefore asserts that karyotype, phenotype, and gender are subtypes of an informational "qualifier" (vs. biological or social) entity.

With respect to marital status, rather than simply state "married" vs. "single", terminologies contain numerous codes that combine other information such as living arrangements (married living apart), marital history (how the most recent marriage ended), stage of life (spinster), length of marriage (newlywed), number of spouses (monogamous), etc. SNCT has 35 marital status codes, including *Eloped*, *Divorced, Monogamous*, *Remarried*, etc. The NCIT has a smaller set of 10 codes, but still conflates status with history (e.g., *Never married*, *Annulled*).

The common practice of such combinatorics with respect to marital status frustrates interoperability – each standard proposes a different set of combinations. An effort to harmonize marital status codes among several standards development organizations (SDOs) such as HL7 and ANSI failed because one SDO went out of business. Characteristic of the combinatory approach, there was subsequent re-opening of discussion and numerous new proposals for different standards and the reconsideration of particular combinations.[5]

3.3 Semantic Web

The Friend of a Friend (FOAF) project is a semantic-web-based effort to standardize information about people (and more). Just as the information model approach merely reifies tables as classes and their fields as attributes (with respect to demographics), FOAF reifies tables as Web Ontology Language (OWL) classes and table fields as relations.[6] For example, it merely replaces the Person table with a Person class, the gender field with a gender relation, and the birth date field with a birthday relation.[7] Worse, the range of the gender relation is the "string" data type. Thus FOAF does not support interoperability of gender data at all, as any string is compliant. The semantic web approach as embodied by FOAF thus does little for our understanding of demographics or for interoperability of demographics.

4 The Realist Approach to Selected Demographics

Our hypothesis in this work was that an ontological analysis of the reality that underlies demographic information could bring coherence to demographic data. We also anticipated a reduction in the need for demographics-specific relations such as "gender", "birthday", etc.

4.1 Dates of Birth and Death

In reality, there is a birth (death) event, and the person who was born (died) is the agent in that event. This event occurs during a particular day. We thus represent John Doe's birth date of February 26, 1981 as:[8]

john_doe_birth **instance-of** Birth[9]

john_doe **agent-of** john_doe_birth at t1

john_doe_birth **occurs** t1

t1 **instance-of** Temporal instant

t2 **instance-of** Temporal interval

[5] As discussed on a thread of the HL7 Vocabulary Workgroup listserv, reproduced here: http://hl7-watch.blogspot.com/2010/11/demographics-hl7-vs-reality-part-1.html

[6] In the semantic web community, a relation is usually referred to as a 'property', after OWL.

[7] The semantic web VCard 'standard' similarly has a birthday property, with a different URI.

[8] Translation of these statements into Referent Tracking templates is straightforward.

[9] Relations between occurrents are not time indexed. See Smith et al. [8].

t1 **during** t2

t2 **denoted-by** '1981-02-26'

We would represent his death similarly. We use representational units (RUs) from the Advancing Clinico-Genomic Trials Master Ontology (ACGT-MO) [9] for *Birth* and *Death*, where they are correctly placed as subtypes of occurrent.

A benefit of this representation is that we can link other information about John Doe's birth (death) to its representation as required. By contrast, the usual approaches cannot accommodate such linking because the dates are just a field/attribute/property about which we can say nothing more.

4.2 Gender

We agree with the WHO that genders are socially constructed roles. We created the Ontology for Medically Related Social Entities (OMRSE)[10] in part to represent gender roles properly.

We then represent John Doe's gender as:

john_doe_gender **instance-of** *Male gender* since t2

john_doe_gender **inheres-in** john_doe since t2

Ceusters and Smith first suggested representing gender as a separate entity, but they did not relate it to the person [10]. We arbitrarily chose John's birth date as the time his gender began to exist. However, John's parents might have learned his sex and thus chosen a male name, purchased male baby clothes, etc. before his birth. We cannot envision a need for a more precise date, but were a better date known and considered important, we could easily use it instead of birth date.

We can also represent transgendered individuals, by stating for example that the person's gender ceased to instantiate *Male gender* and began to instantiate *Female gender* at a given time:

john_doe_gender **instance-of** *Male gender* during t2 to t3

john_doe_gender **instance-of** *Female gender* since t3

4.3 Sex

Phenotypic sex, at the level of granularity of the whole organism, is a quality. The reason is that biological sex refers to more than just reproductive organs. In humans, sex also includes differences in distribution of body hair, levels of hormones, deepness of voice, and so on. Thus, sex differences are widespread throughout the body. The Phenotypic Quality Ontology (PATO) accurately represents *Phenotypic sex* as a descendant of *Organismal quality*.

We represent John Doe's sex using the PATO URI for *Male sex* as:

john_doe_sex **instance-of** *Male sex* since t4

john_doe_sex **inheres-in** john_doe since t4

Sexual differentiation occurs early during fetal development, with notable differences around four weeks. Thus for t4, we could use a date of four weeks after an estimated date of conception. Greater precision is likely not necessary but again, we could use a more precise date if available. Our approach can also accommodate ongoing sexual development by representing, for example, Tanner stages as qualities and tracking instantiation over time as with gender.[11]

4.4 Marital Status

Because it is the legal aspects of marriage that motivate the capture of marital status in EHRs, we focus on its contractual aspects. In the U.S. at least, federal and state governments recognize marriages by conferring upon the couple certain legal obligations and rights. The key rights of concern to healthcare involve hospital visitation and decision making.

Each individual in the marriage is a party to a marriage contract.[12] We represent this role in OMRSE; it is a subtype of *Party to a legal entity*. We then represent John Doe's "married

[10] OMRSE is publicly available at http://code.google.com/p/omrse/through an SVN repository

[11] We note that the usual approaches cannot even begin to cope with this problem, because they do not track sex, gender, etc as single entity.

[12] Common-law marriage implies a contract between the parties that can only be terminated as with other marriages. Even then, only 9 states still recognize it and those states frequently require evidence of mutual agreement before conferring recognition.

status" as (where t5 is the wedding date):

> john_doe_mr_role **instance-of** *Party to a marriage contract* since t5
>
> john_doe_mr_role **inheres-in** john_doe since t5

What about John's unmarried brother Jack? For an affirmative statement that Jack is single, we follow Ceusters et al. [11]:

> jack_doe **lacks** *Party to a marriage contract* w.r.t. **bearer-of** since t6

If t6 is Jack's birth date, then we have successfully captured the semantics of "Single never married" included in many terminologies, but without increasing the number of RUs (or codes) in the ontology.

If either John gets a divorce or is widowed, we can update our representation:

> john_doe_mr_role **instance-of** *Party to a marriage contract* during t5 to t7
>
> john_doe_mr_role **inheres-in** john_doe during t5 to t7

In states that confer on same-sex couples obligations and rights similar to those of marriage, we represent this situation with *Party to a domestic partnership agreement* from OMRSE:

> jim_doe_dp_role **instance-of** *Party to a domestic partnership agreement* since t8
>
> jim_doe_dp_role **inheres-in** jim_doe since t8

This representation also allows us to say other things about the contract. For example, to handle jurisdictional issues, we can represent the relation between the party role and the contract, and between the contract and the jurisdiction that recognizes it in OMRSE. When necessary – e.g., when a marriage in one country is not recognized by another – we can capture jurisdictional information and relate it to the person's role. None of the usual approaches offer this flexibility.

5 The Demographics Application Ontology

Our representation uses RUs from several, different realism-based reference ontologies. To facilitate implementation in RTSs and other applications that manage demographic data, we have created the Demographics Application Ontology (DAO).[13] The DAO does not, and will not, create new RUs for types.[14] It uses the Minimum Information to Reference an External Ontology Term approach [12] to import RUs from reference ontologies.

6 Discussion

Given ubiquity of demographics in information systems used by entities in the economy worldwide and governments, the problem of accurate representation is of critical importance. We have represented the reality underlying key components of demographic data, namely dates of birth/death, gender, sex, and marital status. In doing so, we untangled much of the web of confusion surrounding sex and gender in concept-based approaches. Although Milton was the first to recognize this confusion [13], we have demonstrated its pervasiveness throughout leading concept-based artifacts including HL7, SNCT, LOINC, NCIT, and UMLS. We also reduced marital status to the essence of the rationale for capturing it in the first place. Along the way, we have demonstrated improved flexibility of representation when additional information is required.

Our representation of demographics need not complicate data entry for users of EHRs or other applications. To illustrate this fact, we implemented our approach at http://demappon.info/Demographics.php. It shows a typical, web-based form for entering demographics. After submitting data, the user sees the RT templates that the RTS created behind the scenes, with detailed explanations.

Our approach has the potential to simplify standardization of demographic data. We have removed the requirement for shared field/attribute/property names such as "administrativeGenderCode" (HL7) and "gender" (FOAF). Indeed, nothing in our approach requires demographics-specific relations – we used existing relations from the Relation Ontology (RO) and Ontology of Biomedical Investigations (OBI). Thus, any application that understands these relations in addition to

[13] Available publicly at http://code.google.com/p/demo-app-ontology/

[14] The current version does include fiat classes, a.k.a., attributive collections, as OWL individuals for the purpose of capturing race information. Our work on race is ongoing.

several RUs from PATO, OMRSE, etc. (gathered into the DAO for convenience), will correctly interpret our representations.

Finally, our approach led us to appreciate the diversity of the types of entities involved with demographic data, including qualities (sex), roles (gender and marital status), events (birth and death), etc. that form fundamental distinctions at the top levels of multiple upper ontologies. The distinction between qualities and roles is important because qualities are always exemplified when present whereas roles are not. This distinction is blurred by usual approaches to demographics.

This diversity also illustrates the requirement for application ontologies such as the DAO in the realist approach. The DAO facilitates representing demographics using RUs from reference ontologies, which represent portions of reality without respect to particular use cases. Then, to facilitate a given use case, such as demographics, the application ontology can pull together the needed RUs and relationships among them.

7 Conclusion

Despite the apparent simplicity of demographic data, and thus the expectation that they ought to be easy to standardize, few if any standards for demographics data enjoy widespread adoption. We have illustrated that usual approaches to demographics and standards for them fail to account for the reality underlying them, and that representing this reality has the potential to simplify standardization and increase the flexibility and extensibility of the representation.

Acknowledgements

This work was supported by award numbers 1UL1RR029884 and 3 P20 RR016460-08S1 from the National Center for Research Resources. The content is solely the responsibility of the author and does not necessarily represent the official views of the NCRR or NIH.

References

1. Ceusters W, Smith B. Strategies for referent tracking in electronic health records. J Biomed Inform. 2006 Jun;39(3):362-78.
2. Medicare and Medicaid Programs; Electronic Health Record Incentive Program, 412, 413, 422, and 495 (2010).
3. Scheuermann RH, Ceusters W, Smith B, editors. Toward an ontological treatment of disease and diagnosis. AMIA Summit on Translational Bioinformatics; 2009.
4. Hogan WR. Towards an ontological theory of substance intolerance and hypersensitivity. J Biomed Inform. 2010 Feb 10.
5. Cowell LG, Smith B. Infectious Disease Ontology. In: Sintchenko V, editor. Infectious Disease Informatics: Springer New York; 2010. p. 373-95.
6. Goldfain A, Smith B, Cowell L. Dispositions and the Infectious DIsease Ontology. In: Galton A, Mizoguci R, editors. FOIS. Amsterdam: IOS Press; 2010. p. 400-13.
7. Jorgensen PB, Kjartansdottir KR, Fedder J. Care of women with XY karyotype: a clinical practice guideline. Fertil Steril. 2010 Jun;94(1):105-13.
8. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol. 2005;6(5):R46.
9. Brochhausen M, Spear AD, Cocos C, Weiler G, Martin L, Anguita A, et al. The ACGT Master Ontology and its applications - Towards an ontology-driven cancer research and management system. J Biomed Inform. 2010 May 11.
10. Ceusters W, Smith B. Referent tracking for treatment optimisation in schizophrenic patients: A case study in applying philosophical ontology to diagnostic algorithms. Web Semantics: Science, Services and Agents on the World Wide Web. [doi: DOI: 10.1016/j.websem.2006.05.002]. 2006;4(3):229-36.
11. Ceusters W, Elkin P, Smith B. Negative findings in electronic health records and biomedical ontologies: a realist approach. Int J Med Inform. 2007 Dec;76 Suppl 3:S326-33.
12. Courtot Ml, et al. MIREOT: The minimum information to reference an external ontology term. Applied Ontology. 2011;6(1):23-33.
13. Milton SK. Top-Level Ontology: The Problem with Naturalism. In: Varzi AC, Vieu L, editors. Formal Ontology in Information Systems. Amsterdam: IOS Press; 2004.