The International Journal of Digital Curation
Issue 1, Volume 6 | 2011

Use of Ontologies for Data Integration and Curation

Judith Gelernter,
Carnegie-Mellon University

Michael Lesk,
Rutgers University

Abstract

Data curation includes the goal of facilitating the re-use and combination of datasets, which is often impeded by incompatible data schema. Can we use ontologies to help with data integration? We suggest a semi-automatic process that involves the use of automatic text searching to help identify overlaps in metadata that accompany data schemas, plus human validation of suggested data matches.

Problems include different text used to describe the same concept, different forms of data recording and different organizations of data. Ontologies can help by focussing attention on important words, providing synonyms to assist matching, and indicating in what context words are used. Beyond ontologies, data on the statistical behavior of data can be used to decide which data elements appear to be compatible with which other data elements. When curating data which may have hundreds or even thousands of data labels, semi-automatic assistance with data fusion should be of great help.[1]

[1] This paper is based on the paper given by the authors at the 6th International Digital Curation Conference, December 2010; received December 2010, published March 2011.
The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Metadata and Ontologies in Data Fusion

Data are usually stored in a database, typically a spreadsheet, with element names and values recorded for different subjects, times or situations. These values may be answers to survey questions, or data recorded about a person or an event. Later researchers may wish to combine schemas or information from multiple databases and need to understand which element values may be combined across different databases (Bleiholder & Naumann, 2008; Doan & Halevy, 2004). Even for the familiar problem of geographic location, a database might choose street address, zip code or latitude-longitude, not to mention less common identifiers, such as census tracts or state plane coordinates.

Although databases would ideally choose their labels from the standardized vocabularies that exist in many scientific areas, data collectors are typically not metadata specialists, and what is perhaps most common is a natural language explanation of what a data element is supposed to be. Particularly in the case of a survey, the question asked is often the best, if not the only, definition of a data element's meaning. Local context also matters: a survey asking children if they engaged in "risky behavior" such as "surfing" means something different if it dates from the 1960s.

To fuse data recorded in different ways requires some outside resource that describes the variable names, survey questions or metadata. Ontologies and thesauri are controlled vocabulary lists, often with definitions, sometimes arranged in vocabulary hierarchies to show the relationship among terms. We use the terms interchangeably here. Ontologies have been used successfully to mediate between schemas (Hu & Qu, 2007, Bouquet, 2007). Difficulties increase when more than one ontology is necessary and the ontologies themselves require alignment (Cruz et al., 2009), although crosswalks sometimes exist. Furthermore, schemas and ontologies evolve, so the mappings between schemas need updating (An & Topaloglou, 2008).

As an example of graphical presentation that may assist in data integration, Figure 1 excerpts a part of two dental ontologies and shows links between them, in this case "floss". Each record it as part of the patient's history but organize the field in different ways (both ontologies contain many more categories than shown).

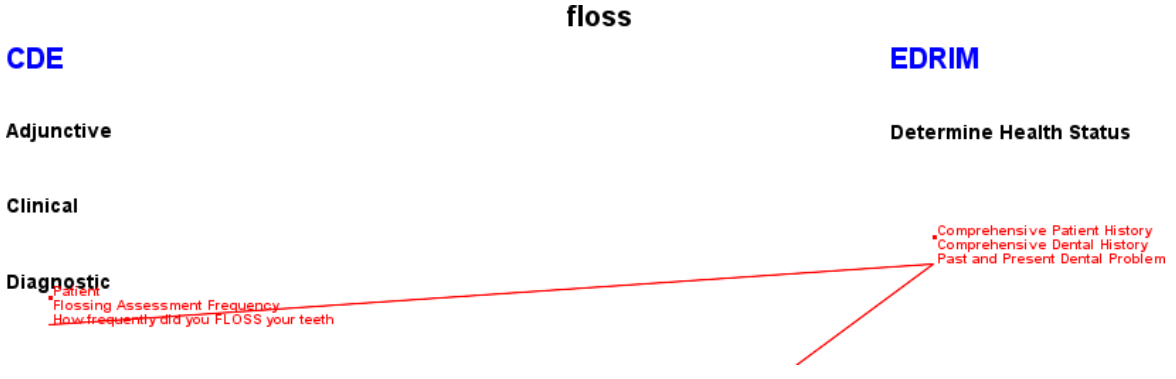

Figure 1. Comparison of CDE and EDRIM Dental Ontologies.

We have worked with multiple subject areas, including surveys on addiction (Tarter & Vanyukov, 2002), on aging (NACDA, 2010), and on the description of dentistry (CDE, 2010). Ontologies can assist in at least semi-automation of data fusion. Our software tools try to identify probably-similar items by counting word overlaps, using ontologies to help weight the importance of words and to detect synonyms. It then asks human curators to validate the results and also to approve suggested scaling based on observed ranges and distributions.

Same World View among Database Schemas

In some cases, the basic data collected is roughly the same. Two different questionnaires about the elderly asked respondents if they suffer from any of a list of health problems including "diabetes" or "heart trouble". Those two entries were asked exactly the same way in both surveys, and so can be easily matched automatically. However, one survey asked about "high blood pressure" and the other about "blood pressure (abnormal)." One asks about "breathing difficulties" and the other about "asthma" or "emphysema or chronic bronchitis." And one says "digestive problems" while the other divides them into "ulcers" and "other stomach or intestinal disorders."

In general, word matching using a general purpose, non-domain thesaurus can be used to detect many of these overlaps. Even the very general *wordnet* (Miller, 1995) or *moby* thesauri (Ward, 2000) will relate "asthma" to "respiratory" and to "breathe". But the cases where a single condition is subdivided or merged present more of a problem. Even though the two surveys are the same "in the large", there may thus be detailed examples of one-to-many or many-to-one mappings. In general, one can imagine amalgamating the data in a many-to-one mapping to get the simpler form. Dividing the more general form to decide whether somebody who has only reported "respiratory" problems might specifically have "emphysema" will usually be impossible.

Our experimental interface searches through collections of survey text for apparently matching variable names or questions. For example, Figure 2 shows a relation detected between two variables whose names are somewhat coded (*oppositional-defiant-disorder* vs. *indirect-hostility*). We use the question that supplied the variable data to automatically find a match on the word *temper*.

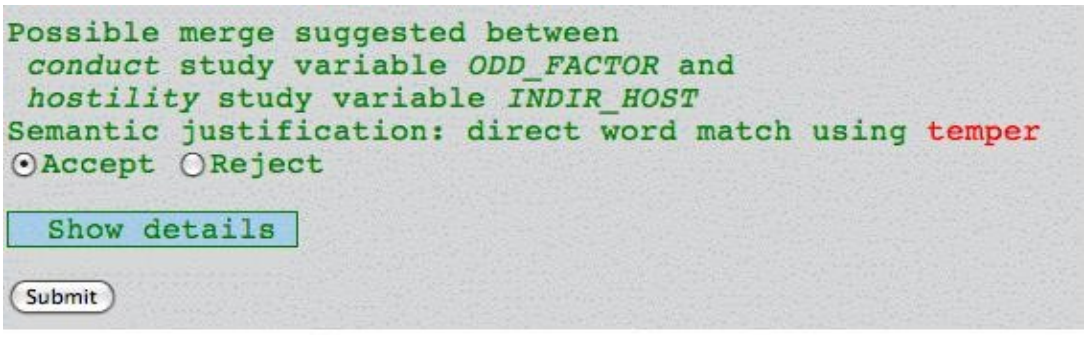

Figure 2. Potential Data Element Identity.

Surveys may request answers as binary, categorical or scaled. Compare:

Do you have trouble breathing?	Yes / No
Do you have trouble breathing?	None / Rarely / Sometimes / Often / Always
Please rate your breathing on a 1 to 5 scale:	(5 is Easiest)

Sometimes actual numbers are requested: questions about commuting or weight usually expect an answer in miles or pounds. We might be able to find similar questions, but will still need to convert answers to the lowest common data form (both to nominal, or both to ordinal, etc).

Fusion can require more complex calculations as well. "Age at school leaving" is related to "highest grade level achieved" but usually involves adding 7. Even worse is "age" compared with "year of birth"; not only must one adjust by the current year, but the numbers move in opposite directions.

Can this be done automatically by looking at the range and distribution of the answers to the questions? Suppose we have two questions that are the same, and one has answers ranging from 1 to 3 and the other from 1 to 7. This can easily happen if one question uses a Likert scale and the other uses words:

Please rate your walking ability:	Poor / OK / Good
Please rate your walking ability from 1 to 7	7 is Good.

If we have enough data to reliably measure the distributions of the answers, it is possible to suggest that the second number could just be divided by 2.3 to give comparable values. Our prototype has an option to display a histogram of values, as shown in Figure 3, to help the user decide on ranges and central points. If it believes two different data elements should be fused and are comparable, it will suggest a scale transformation for the curator to consider.

If we have extremely good statistical properties, we can imagine filling in missing data on the basis of prediction from data we already have, but only the rare survey has such well behaved data (Langford et al., 2001).

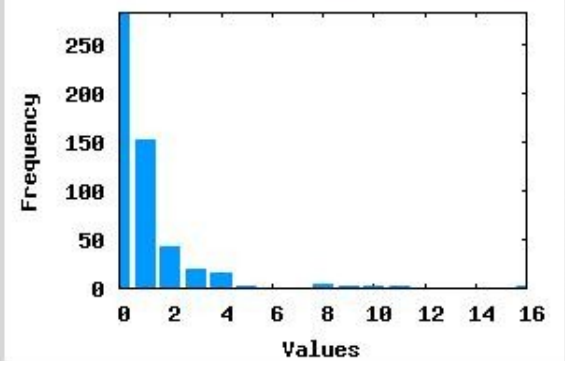

Figure 3. Histogram of Answer Values Observed for One Query.

Unfortunately, there are many practical problems doing this. For example, many surveys are only looking for a few people that have some problem. If 90% of the people answering your survey rate their walking ability as "good," and half the remainder did not answer the question, there may not be enough data points to feel confident about suggesting automatic scaling.

Curiously, in a few cases evenly spread data may pose a problem. Suppose you have two queries on the same subject, but with different answer coding. If the query asks: "Have you ever been arrested?" with replies coded either "Y/N" or "1/0", but the answers are 90% one way and 10% the other, you can guess that the two 90% replies should be matched. But if the question is whether you are a man or a woman, and one survey codes the answer "M/F" and the other "0/1", since both answers are likely to appear 50% of the time there is no way to guess the alignment without other information.

We tried a pilot experiment involving people looking at the results and either approving or disapproving the computer-suggested question alignments in drug addiction data. Participants essentially acted as data curators helping to integrate given data sets.

Subject groups were asked to find matches manually: one group found matches among match-rich data sets (75 variables total) and the other group found matches among match-poor data sets (75 variables total). Test groups from each had the advantage of the use of our prototype, which uses the variable name and the question that elicited the variable to find matches. They could also see a histogram of the distribution of data values in each survey, as shown in Figure 3, to aid in their decisions.

Not surprisingly, participants without the prototype system to help them find matches took much longer: 37 minutes on average rather than 9 minutes. We used one dataset carefully chosen to have many matching variables and one dataset chosen with fewer overlapping variables. In the data set with many matches, the subjects found 3.8 pairs per minute with the software and 1.5 without. In the dataset with few matches, subjects found 0.32 matching pairs per minute with the software and 0.19 pairs without. However, with only four subjects in the pilot experiment, the results in matches per minute were not significant ($p=0.10$). More important for us is that the average accuracy remained high, although again, the small number of subjects and the high scatter does not yield a significant result. The number of incorrect matches found is shown in Figure 4, with the red bars indicating subjects working without the software and the blue bars showing the subjects with help.

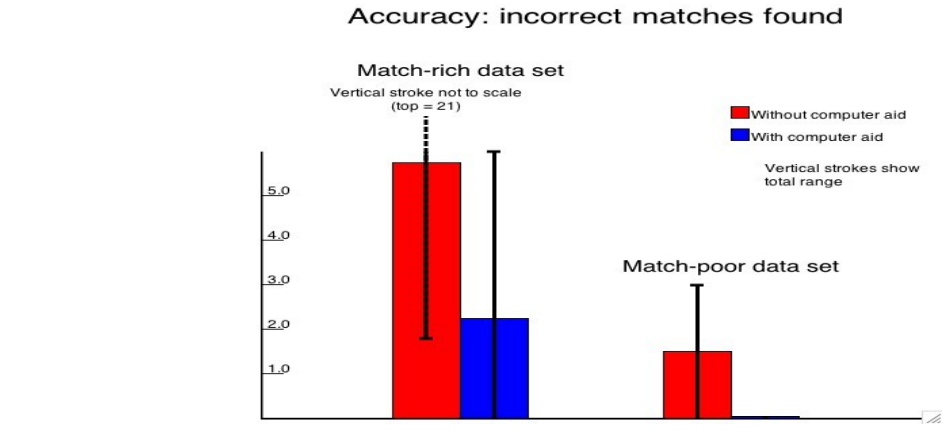

Figure 4. Accuracy of Data Integration With and Without Software Assistance.

Different World View among Database Schemas

Sometimes data schema, even when dealing with the same subject, have a completely different view of the world. Our examples in this section come from the EDRIM schema (Acharya et al., 2009), which is a practical representation of dentistry and the Common Data Element schema from the National Institute of Dental and Craniofacial Research (NIDCR) (CDE, 2010). In general, EDRIM is oriented around what a dentist sees or does. We also use MeSH (MeSH, 2010), which is both more general (since it covers all of health and medicine) but also more limited, since it does not cover aspects of practice that are unrelated to research or basic knowledge. For example, the EDRIM data element "patient's next appointment" does not appear in MeSH.

Of course, not all of these data elements will be collected for each patient record. Aside from the obvious (no dentist, even using EDRIM, will ask male patients if they are pregnant), there are nearly 1000 concepts in EDRIM. If a dentist really went through asking all patients the date of their last blood transfusion, every dental examination would take forever. Again, this implies missing data frustrating some statistical methods.

We use word overlaps to identify matching elements across data schema, and use ontologies to bridge data sets and schemas. Ontologies can:

1. Identify which words should be considered important for comparisons.

   For example, the NIDCR data schema has 11 elements in which the phrase "a person who requires medical care occurring before the period of time that it takes for Earth to make a complete revolution around the sun, approximately 365 days" appears. It is unclear why the creators of this data system thought it necessary to define the ordinary word *year*. However, to avoid thinking this is significant, we can use only words from the dental categories in MeSH.

2. Identify synonyms.

   For example, EDRIM uses *skin cancer* where CDE uses *carcinoma* and MeSH prefers *skin neoplasm,* but all of these are linked in the NCI thesaurus.

3. Identify hierarchical level.

   Concepts that may be at the top level for one ontology may be low down for another. For example, EDRIM's dentist's-eye view puts patient afflictions underneath diagnosis, whereas MeSH puts diseases at the top of the hierarchy.

In short, we use ontologies to bridge different data schemas and the data that fill those schemas. We show this graphically below. The disease-centered model of CDE presents "implant" as something in the patient history, or as something being done right now, or as a preventive action. In the MeSH ontology bridge, it is a material (something used in treatment) as well as a procedure; EDRIM uses it only once.

Figure 5. Simplified Diagram of Occurrences of "Periodontal" Terms in Three Dental Ontologies.

Conclusions

Future data curation will need tools that enable re-use of data that are gathered at different times by different people. Ontologies can assist data integration by mediating between schemas. Even though these ontologies are applied automatically for information retrieval, we still see a need for curators to validate the results.

Acknowledgements

Our thanks to Professors: Levent Kirisci, Ty Ridenour, Titus Schleyer, Ralph Tarter and Michael Vanyukov of the University of Pittsburgh for access to surveys and ontologies.

References

Acharya, A., Mital, D.P., & Schleyer, T.K. (2009). Electronic dental record information model. *International Journal of Medical Engineering and Informatics, 1, (4).* Geneva, Switzerland: Inderscience Publishers.

An, Y. & Topaloglou, T. (2008). Maintaining semantic mappings between database schemas and ontologies. *Semantic Web, Ontologies and Databases: Lecture Notes in Computer Science, 5005.* Heidelberg and New York: Springer-Verlag.

Bleiholder, J. & Naumann, F. (2008). Data fusion. *Computing Surveys, 41, (1)*. New York, NY: Association for Computing Machinery.

Bouquet, P. (2007). Contexts and ontologies in schema matching. *Proceedings of Context and Ontology Representation and Reasoning.* Roskilde University, Denmark.

CDE Browser. (2010). Retrieved January 10, 2011, from https://cdebrowser.nci.nih.gov/CDEBrowser/.

Cruz, I.F., Antonellia, F.P., Stroe, C. (2009). AgreementMaker: Efficient matching for real-world schemas and ontologies. *Proceedings of the Very Large Database Endowment 2, (2)*. Lyon, France.

Doan, A. & Halevy, A.Y. (2004). Semantic integration research in the database community: a brief survey. *AI Magazine, 26, (1)*. Menlo Park, CA: American Association for Artificial Intelligence.

Hu, W. & Qu, Y. (2007). Discovering simple mappings between relational database schemas and ontologies. *Lecture Notes in Computer Science, 4825*. Heidelberg and New York: Springer-Verlag.

Langford, E., Schwertman, N. & Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician, 55, (4)*. Alexandria, Virginia: American Statistical Association.

Medical Subject Headings. (2010). *United States National Library of Medicine.* Retrieved January 10, 2011 from http://www.nlm.nih.gov/mesh/meshhome.html.

Miller, G.A. (1995). Wordnet: A lexical database for English. *Communications of the ACM, 38, (11)*. New York, NY: Association for Computing Machinery.

National Archive of Computerized Data on Aging. (2010). *Changing Lives of Older Couples* (Study 3370); *Survey of Low Income Aged and Disabled, 1973-1974* (Study 7661); *Chinese Longitudinal Health Longevity Survey* (Study 24901); *Aging, Status and Sense of Control* (study 3334); and *Aging in the Eighties* (Study 8691). Retrieved January, 10, 2011, from http://www.icpsr.umich.edu/icpsrweb/NACDA/.

Tarter, R. & Vanyukov, M. (2002). *Etiology of Substance Abuse Disorder in Children and Adolescents: Emerging Findings from the Center for Education and Drug Abuse Research* . Binghamton, NY: Haworth Press.

Ward, G. (2000). Moby. Retrieved January, 10, 2011, from http://icon.shef.ac.uk/Moby/.