Reviewer's report

Title: Text data extraction for a prospective, research-focused data mart: implementation and validation

Version: 1 Date: 17 April 2012

Reviewer: Joanne Luciano

Reviewer's report:

Please note - I do not run SQL and therefore could not evaluate the software that the paper was about. The comments below are limited to the paper.

Major Compulsory Revisions

1. I do not generally work with in Enterprise Data Warehouse (EDW) environment, but I do have extensive experience and knowledge in the area of computer science and translational research. As a reader of this sort, I would be better by a decreased use of Three Letter Acronyms (TLA). Spelling out, defining terms such as data warehouse and data mart, and how they are different from the more familiar database would improve understanding of the paper and make it easier and more enjoyable to read. For example, that a data warehouse is optimized for querying the data while a database is optimized to record data and will be confused by this. The same is true for EHRs, chart review etc. Adding a few clarification sentences will greatly improve the readability and accessibility of the manuscript because they will give the reader a more solid grounding in the concepts introduced, which will support the acronyms are used throughout. For example, the title of the paper is “Text data extraction for a prospective, research-focused data mart: implementation and validation” but the term data mart is never defined, nor is any prose given to provide the reader with an appreciation of what it is and why its useful, and how it relates to a data warehouse. Furthermore, it would be helpful to include with the definition of terms, examples of what they mean and why they are useful and important in this context.

2. It would be beneficial if there were a way to try tool outside of the SSIS (SQL Server Integration Services) or SSRS (SQL Server Reporting Services) environment (perhaps with a demo) to reproduce the results, or otherwise explore the software in order to determine its utility. I can imagine a web-based service that performs the regex parsing could be useful to a broader community.

3. The example of the Pulmonary Function Test needs to be fleshed out more. For example,

“Pulmonary function tests (PFT) [a] are administered using a SensorMedics Vmax Encore PFT Autobox Pro machine [b] and data is captured using SensorMedics software version IVS-0101- 21-1A (both CareFusion Corporation, San Diego, CA). PFT data is stored in a proprietary [5] database used by the
SensorMedics software.

[a] It would improve the paper to provide a description of what the PFT test specifically measures (how is “function assessed”), why the test is performed, and what the data measurements obtained tell me, would make this sentence interesting and informative]

[b] SensorMedics Vmax Encore PFT Autobox Pro machine?? a what? I am sure the reader will understand if you would give them a chance – I know that it doesn’t matter, and the only thing that is important for your software is that the instrument generates “data”, but I would like to learn something about the example too, since you are illustrating it.

Therefore, a scenario would be useful to help the reader understand the software, by example, and in context, the problem that software solves and how it solves it. The section on clinical utility could provide a bit more – something along the lines of what’s in Wikipedia (“is a systemic autoimmune disease or systemic connective tissue disease that is a subtype of scleroderma (a chronic systemic autoimmune disease (primarily of the skin) characterized by fibrosis (or hardening), vascular alterations, and autoantibodies). [2] It is characterized by deposition of collagen in the skin and, less commonly, in the kidneys, heart, lungs & stomach. Female to male ratio is 4:1. The peak age of onset is between 30-50 years.” And maybe a little about the prevalence So I learn something and to warrant the paper being in a BMC journal rather than an ACM. That being said, it is commendable that the scleroderma investigators have been able to build additional queries and use for other research projects.

4. Figure 1 is good. An excerpt of it should be used to walk the reader through an example. Show how Table 1 and Figure one relate.

5. Figure 3 is unreadable - the font is too small. Also, there is no annotation. In the text show a the text and the regular expression that matches the pattern.

6. Table 1 could use more detail in the description. As it stands, I do not get much out of it.

7. Would Figure 3 be better represented as a graphic with indicating a cycle since it occurs every day (or am I reading it wrong?) It would also be more informative to indicate where on this diagram the software tool is operating. I am guessing that it is in step going into to the EWD Discrete PFT data mart (and is the data mart within the EWD?) A graphic would be helpful here.

Minor Essential Revisions

1. Cut down on the use of three letter acronyms. It’s not that big a paper or that cumbersome, especially for the terms that are used less frequently, such as ODS. OK to put it in parenthesis after the initial use, but better to not use the TLA.

2. In the Technical Generalizability of Text Processing Workflow section, please show or describe how “The EDW makes it easy to construct regular expression parsing pipelines for text.”
3. RE: Downloaded 315 times since 6/15/2011 – that date is nearly a year old. I'm wondering why? ....

Discretionary Revisions
1. I'm also wondering if this technology is “legacy” and how it compares to or fits in with other translational research pipelines such as i2b2 or if the output can be adjusted to generate linked open data?
2. I'm interested (as a reader) in knowing what the details of 6 inconsistencies were – what were the mistakes in the manual abstraction (and why, especially since the chart abstractor was familiar with PFT)?

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.