

## A Place For Everything, and Everything in It's Place

By Paul Curry

(slightly modified by Alan G. Labouseur)

*Guten Tag. Ich heisse Paul. Wie geht es Ihnen?* No, this isn't for a language class, but rather to demonstrate the importance of context in determining the information value of data. The letters of the English alphabet are symbols we use every day, as are Arabic numerals (1, 2, 3, etc.). Because they are in constant use, we begin to feel they have inherent meaning; further, certain groupings of letters, such as 'Tag' above, are commonly used words in English. However, the same letters, assembled in the context of a different language, can carry completely different meanings. The layout of the three sentences above indicate, to English speakers, that they *are* sentences, but besides 'Tag' and 'Paul', they may seem completely without meaning. In German, they mean (roughly) "Good day. My name is Paul. How are you?". Similarly, the poem "Jabberwocky" demonstrates that otherwise meaningless symbol-strings can be imbued with meaning by their context. The essential need for context is a feature of all human attempts at communication, and examples therefore abound: written language, spoken homonyms, musical scores, art, and so on *ad infinitum*.

In databases, we store enormous collections of raw data, which we hope embodies information we find interesting. At the lowest level, the machine on which the database runs does not know what we are trying to store: all it comprehends are bit-fields of specified lengths. The database is subtler: it knows that we are storing character strings, integers, floating-point numbers, or arbitrary binary objects. At the metadata level, it knows about relations between some of the data, in that it knows what data is stored on what tables. However, only viewing the data through a context delivers up information.

The Oberlin College student records database easily demonstrates this property of databases. The raw data input may very well look like "CURRYPaulR90045688607498123521655557891313MooreLanePeekskillNY105663.09CSCI2103.67MATH3992.33MATH2013.67PHIL260NE" - and so on. It is a string of alphanumeric characters. Through the eyes of the default American context, and possessing what information we already have, some of it seems to make sense. There are

what *looks* like a name, an address, and some course titles. There are also things for which we have no context – the long string of numbers between H and S, for example.

By providing more structure to the data, we can deduce more from it. Thus, the string could become “{((CURRY, Paul, R), 900-45-6886, 321-21-1234), ((1313 Moore Lane, Peekskill, NY, 10566), 216-555-5789), (3.09, (CSCI210, 3.67), (MATH399, 2.33), ...), ...}”. The address becomes distinct, as does the name; course grades, a phone number, and overall GPA emerge. We see two 9-digit numbers, grouped similar to a Social Security Number, but we can’t immediately determine whether either number is in fact an SSN, or which one belongs to this student; we require still further structure and context to extract that information from the data.

It happens that all Oberlin student ID numbers in the relevant time period were 9-digit, and formatted like social security numbers. Thus, it would be a reasonable assumption to assume that 900-45-6886 was a student ID number, belonging to this student; it would in fact be a correct assumption. That information could only be extracted from the data, however, because the data was presented in a sufficiently structured format and certain properties of the context were known to the inspector. Better still; had we been presented with labeled columns of data, we wouldn’t have had to deduce the intended meaning from its structure.

Without such context, the raw character string would indicate no more than it did on first inspection, and that only to someone who knew it was an American college database. To a native Chinese speaker, for example, the alphanumeric symbols might be entirely without context, thus signifying nothing.

In all attempts to record or convey information, the success of said conveyance is predicated not solely on the receiver correctly receiving the message (the symbols), but also in their understanding of the medium in which the message has meaning (the context). Without the message, there is no communication, but without grasping the medium, there is **mis**communication, which is potentially far more dangerous. Without both, therefore, communication has failed, and all the character strings in the world will behave one nothing, without a Rosetta Stone of context.