OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

# The Impossibility of Global Consistency

**GIO WIEDERHOLD**

IT APPEARS THAT THE ISSUE OF MUTUAL GLOBAL CONSISTENCY is becoming a major research challenge for biological data management. Problems due to inconsistencies arise whenever we try to move from proven solutions in local or delimited information systems to larger scale systems.

Our science depends on words to express our ideas. Even when we replace words by codes, as in clinical diagnoses, words remain the medium in which we share the meaning, the semantics, of the concepts being discussed. Words derive their power in science from identifying conceptual abstractions. When we are working in a narrow context, we have a very clear mapping of a concept to its instances, say a set of patients that have been diagnosed with a specific disease. We link the specific disease with a set of symptoms and signs. Higher level abstractions group diseases into categories, at ever higher levels. Now the mappings to actual instances, the metric of consistency, become hard to follow. Furthermore, there will not be a single hierarchy. Each finding and observation is described by many attributes. All these attributes define concepts in distinct hierarchies at multiple hierarchical levels. Still, within one domain consistency will be high.

## INTEGRATION

Our efforts in science must reach out, and be broad. In another clinic, with another set of experts, and a new set of patient instances new mappings will be used from their concepts to instances. We experienced many years ago that data from our tertiary care setting could not be merged with data from private clinics, nor with data from Canadian settings where healthcare was well-nigh free. The patient distributions differed so much that diagnostic terms were applied differently, even though all settings treated patients in the same disease categories. A major effort, multi-year effort to standardize terminology helped, but even after years of sharing data many inconsistencies remained.

When we interact with domains that have fundamentally different objectives, inconsistencies among terms become even more pronounced. A researcher will use terms differently than a clinician, a surgeon will use terms with a more definite emphasis than a pathologist. Similar distinctions occur in all fields of biology. Since our knowledge of biology is increasing rapidly, and is less regulated than medical care, terminological divergence is greater. having descriptive information, metadata, helps in understanding similarities and differences.

We recognize three conceptual levels of integration:

1. *Data integration:* Data from distinct sources are brought together into a single result. The meaning of the data is implicitly assumed to be the same. Metadata, if any, is used passively. All value encodings, numerical and textual, are identical.
2. *Information integration:* There is explicit metadata. Where the semantics do not match transformations are applied. The focus is on the data that are needed to produce information, and value encodings are

Department of Computer Science, Medicine and Electrical Engineering, Stanford University, Stanford, California.

transformed to match the needs of the objective. Where the abstraction levels differ, data are aggregated before merging. The metadata are transformed as well.

3. *Process integration:* Information is extracted using analysis programs within their source context. The results may require still information integration. New metadata are created to describe the result.

The relative volume of the result versus the sources is reduced in each of these three levels. Having a recoded workflow description helps in updating results when the sources change. The semantics, and their metadata representation are handled differently in the three cases.

## DEALING WITH SEMANTICS

The obvious solution is to define each term as unambiguously as possible. But it is rare that biological concepts can be completely specified with measurable parameters, so that dependence on words remains. Examples can help. If a concept defines a certain set of actual patients, or cells, or mice, then consistency can be tested and perhaps achieved. But consistency grounded on extant instances disables generalization and projection into new populations, the objective of science. We see terminological mismatches in many fields, especially in topics as genomics, where rapid growth in many subfields has cause term usage to diverge.

Establishing a committee with members from diverse groups to establish terminological consistency is unlikely to help. First of all, a committee is a mechanism to define compromises, and compromises are likely to create less precision, hurting all participants. It is also difficult to enforce acceptance. A financial stick can help, as in the case of reporting diagnoses and related treatments to insurance companies, but now reimbursement considerations will cause misuse of terms and reduce truthfulness.

Enforcing consistent naming for disjoint enterprises will also make those enterprises less efficient. If say, seven terms are used to partition one concept into seven groups, then efficiency of language is high if those groups are approximately equal in size. But in a different setting dealing with another population, the groups may be very unequal in size, so that word usage becomes inefficient. A common case is the use of abbreviations. In any setting abbreviations are used to efficiently refer to frequently occurring cases. Outsider are often bewildered when insider lingo is full of such usage abbreviations. But when the same insiders enter the outsider's domain, they will be similarly bewildered.

In summary, the words we use to denote concept are precise only in their narrow domain. The mathematical notion that $a = a$ independent of setting fails. It is wise to recognize that we cannot expect full consistency, and instead try to deal with inconsistencies.

What are methods to deal with inconsistency? We will consider formal mappings, applying domain corrections, independent processing, and the explicit introduction of additional uncertainties.

1. If we can formalize the sets of terms in a domain, that is, establish an ontology, then we can proceed with a defining a mapping for interoperation. Such mapping will define when the same word has the same meaning in two domains, when there are synonyms in the two domains, when terms are used in a narrower or broader sense, and when homonyms are completely distinct. It is not necessary that all terms from two domains be matched, we only have to define the mappings for terms used to link information from two domains, with respect to some objective. We call such a linking set of terms an articulation.
2. If data have to be merged, then corrections may be applied. Adjustments may be applied to make the distributions equal. Such adjustments have been used in survival statistics, to account for population differences.
3. If we have similar data from disjoint populations, we can keep that data, and perform the same analyses on all the datasets. The results will require fewer concepts, which can all be reported, and when compared, take the source data differences into account. Such reporting places an additional load on the reader, but will be closer to the truth, and can prevent results from one distribution to be applied to a different one.

4. If no corrections seem feasible, then it simply must be recognized that the number of degrees of freedom increases with every set of data that comes from a distinct source, weakening the confidence in the result.

We phrased our solutions primarily in statistical terms, because statisticians have had to deal with these issues for a long time. However, many analyses today are not statistically based. We encounter equivalent problems in information retrieval, where keyword searches will return incomplete and imprecise (i.e., wrong) results due to inconsistencies of term usage. When searches are broadened, using thesauri or tools as UMLS, completeness is improved, bur precision is reduced.

## CODA

In the end, we will have to live with inconsistencies. We cannot wish it away. We cannot demand that our collaborators use and adhere strictly to definitions that we establish. We cannot expect that a global language police will assure that we all use words so precisely that mutual inconsistency disappears. We just have to take care that inconsistencies among domains does cause us to make serious errors.

## POSTSCRIPT ADDED AS A RESULT OF DISCUSSION: WHY NEGOTIATION CANNOT RESOLVE INTRINSIC SEMANTIC DIFFERENCES

Many people, when first encountering semantic differences, assume that such differences can be resolved by negotiation. While it is fine when such an approach is feasible, it often is not. If negotiation is used in cases where it should not be used, compromises are made that lead to imprecision, resentment, or both.

I will use simple examples to make my point for several types of semantic differences:

1. *Differences in granularity:* When I, as a homeowner, work with my son in fixing the house and need a nail, I will ask for a yeh-long thin nail, with a big head, and he will find what I need in the coffee can we use for various nails. A carpenter has nearly as many terms for nails as the apocryphal Eskimos have for snow: sinkers, box nails, etc. It would be inefficient for the carpenter to use our terminology, and impossible for me to adopt the carpenter's, and the plumber's, etc. terminologies in my household chores.
2. *Differences in scope:* I have observed that in all but the smallest organizations payroll and personnel files are distinct, and managed by different departments. That means that soon the definition of a key term, as "Employee" will differ as well. As a retired faculty member I will not be on the payroll, but I insist on staying in the personnel file so that the right to my little emeritus office and free parking is documented. Some students that receive stipends from some donor fund will be on the payroll, without being part of personnel. It is better to understand and account for the differences than to force these files, and the programs that operate on the files to keep their entries and processing decisions in lockstep.
3. *Temporal differences.* Cash accounts should always be current. For quarterly analyses one will want to have stability at the end of the quarter so that various analyses can be run and compared. Synchronization of what appears to be identical data is again unwise.

All these differences are of little import until integration is attempted. Negotiation is futile. Starting with the assumption that terms from distinct domain are different, and then developing matching rules only among elements that are needed for their articulation to solve some new application is adequate, and fosters cooperation rather than resentment.

I encountered this type of problem more than 20 years ago, before I was comfortable with terms such as semantics, articulation, or ontology. When designing database schemas for hospital information systems we had committees of all the concerned participants: physicians, nurses, pharmacists, billing clerks, etc. The compromises that emerged, not surprisingly, favored the physician's view. When the systems were implemented, the nurses, performing most of the data entry, found the systems awkward, and complained bitterly. The solution was to hire additional staff for data entry, negating the expected cost sav-

ings that provided the motivation for installing the system. I expect that today such an error will not be repeated.

Address reprint requests to:
*Dr. Gio Wiederhold*
*Department of Computer Science, Medicine*
*and Electrical Engineering*
*Stanford University*
*Gates Computer Science 4A*
*Stanford, CA 94305*

*E-mail:* gio@cs.Stanford.edu