# Toward a grammar of data

Alan Freeman 27 September 2020

This note aims to provide a rationale for the classification systems in the Macroeconomic History project. This project, which has a long provenance, tries to meet the practical need to make economic data accessible to, and usable by, the general public. It originates, however, in a problem I encountered earlier, in my work as a practicing economist: namely, everyone that supplies economic data follows different rules. Moreover these rules are not only quite arbitrary, but lead to quite serious mistakes (see for example Freeman 2020)

The idea behind the microhistory project is that economic data, from a variety of sources, should be presented in a commonway. The project has made a first step by introducing common *naming* conventions – to overcome the problems arising because, for example, the country which the IMF calls ‘CONGO, DEM. REP. OF’, UNCTAD calls ‘Dem. Rep. of the Congo’, the UN calls ‘D.R. of the Congo’ and the World Bank calls ‘Congo, Dem. Rep.’

However the need for some kind of common framework extends beyond naming, to the *classification* of data, that is, to some notion of its structure. In the cited article I showed that, for example, the extend and causes of inequality between nations depends crucially on which nations are classified as members of the global South, and which the global North.

The problem is not just that the existing classifications are ‘wrong’ but that data providers do not even treat classification as a problem – it is either supposed they may impose any structure they like, or that this structure is self-evident, as when they speak for example of ‘North America’, without realising that it makes an enormous difference to economic analysis whether Mexico is included in, or excluded from, this category.

The problem is not solved, however, by simply settling on one particular classification and declaring it to be ‘correct’. This is because a variety of different classifications are possible, and the conclusions that can be drawn about the world economy are vastly different, depending on which is adopted. What is required is, therefore, a systematic way of describing these differences in order that the descriptions used can enter into the framing of the research to be conducted.

This way of thinking about classification, however, radically changes the perspective. The problem is not standardisation, which is the usual way of thinking about metadata; rather, it is one of creating a *language for talking about metadata*.

But it should be no surprise either to linguists or to data scientists that as soon as one speaks of language, one is obliged to confront the twin issues of *semantics* or meaning, and *syntax* or grammar.

The *semantics* of talking about countries and regions, to give a simple example, concern such issues as ‘what do we actually mean when we speak of Europe’ – or of ‘poor’ or ‘indebted’ countries, not to mention such charged notions as ‘rogue’, ‘failed’, or ‘aggressor’ countries.

The *syntax* arises because the classification problem presents itself, in the first instance, as one of constructing lists and hierarchies, for example of containment or membership. When we say that France is a ‘European’ country for example, we mean that it is geographically contained in the boundaries of Europe. And when we say that Brussels is a European city, we mean both that it is in the boundaries of Europe and that it is in the boundary of a European country.

Hierarchies also intersect; notably, as do geography and time. ‘Europe’ at the outbreak of the first world war was a vastly different system of polities to what it had became by, say, 1924. A way of talking about this intersections is also required

In the Macrohistory project we have addressed a number of the problems of geography, as manifested in the structures of the geography dimension file. We have, however, only scratched the surface of the classification of economic *indicators,* and have recently undertaken a major re-organisation of the indicators dimension, which this short piece attempts to explain and motivate.

TBC

## References

Freeman, A. 2019. [Divergence, Bigger Time The unexplained persistence, growth, and scale of postwar international inequality](https://www.academia.edu/39074969/Divergence_Bigger_Time_The_unexplained_persistence_growth_and_scale_of_postwar_international_inequality)