

# **A Modest Proposal for a Universal, Object-Oriented, Hierarchical, Extensible Naming System for Population Census Information**

Presented to the 19th Population Census Conference,  
Beijing, China, 26-28 April 2000

**Griffith Feeney**

April 2000

## **Abstract**

A systematic way of naming population census data resources is briefly and impressionistically sketched. The objective is to facilitate (i) the organization and management of census data processing by the census organization and (ii) the dissemination and analysis of census data within and beyond the census organization by (iii) providing easy, rapid, and potentially *automated* access to census information over computer networks. The proposal draws on various developments in information technology, including the Internet domain name system (DNS), hierarchical file systems, relational data bases, and the world wide web. The ideas are general and might be extended to many other kinds of data.

## **A modest proposal for *what?*!**

I'll define those fancy looking words in the title shortly, but first something about the motivation.

Developments in information technology in general, and in computer networking in particular, hold out the possibility of radically improving the way we do our work. Suitably designed network tools will provide easy, rapid, automated access to all information about a population census, from planning documents to editing specifications and tabulation plans to final tabulations, microdata samples and publications. This technology is important within the census organization as a tool for managing census operations and as a way of disseminating census information to users.

The potential of computer networking is best exemplified by the emergence and extraordinary growth of the World Wide Web (Berners-Lee 1999). World Wide Web protocols are designed to accommodate *any* kind of information, and in particular to escape the hierarchical organization imposed by more traditional information retrieval mechanisms. Population census data involves highly specialized formats and is rigidly hierarchical. There is accordingly a poor match between population census information and World Wide Web protocols.

Carrying out a population census involves the creation of a very large number of computer files, not only of "data" proper, but planning documents, enumerator instructions, pretest results, coding protocols, editing specifications, tabulation plans, publication schedules, press releases, and so on. Imagine all these documents and data files organized into a single, large, virtual file system, and suppose that this file system is created at an early stage in the work on a particular census. At the outset this file system will provide "a place for everything", but nothing will be in its place. A

useful view of the work of census processing is the gradual filling up of this file system with the required files. When the last file is added, the work is done!

The person in overall charge of census operations would be able to “drop in” any location in this file system and see the current status of the work. When was this work scheduled to be completed? Is it completed? On schedule? It should be emphasized that what the manager sees is not merely *reports* on the work, but the work itself, “in real time”. Persons in charge of particular parts of the census operation would have similar access to the files for which they are responsible, and perhaps, for improved coordination, to the files of related areas, or even the entire file system.

When work on the census is complete, the networked file system of census information could be put to use in the dissemination of data to users. Appropriate portions of the system would be copied from the internal network to (most likely) a computer connected to the Internet. Metadata—the census schedule, editing specifications, tabulation plan, classification protocols, publication lists, and so on—would probably be provided at no charge on the grounds that readily available metadata is the best way to encourage use of the census.

Nearly all national statistical offices have websites already, of course, but anyone who has tried to collect census information from these sites knows how difficult it is to do so. To realize the potential of the Internet for information sharing it is essential to structure the metadata and data so that searches can be *automated*, that is, carried out automatically by computer programs (sometimes called “spiders” or “robots”) that systematically visit pertinent internet sites, look for the desired information, acquire what is available, and present the results of this data gathering enterprise to the user in “ready to use” form.

A librarian in Topeka, Kansas, for example, (or anywhere else in the world) should be able to issue a single, simple query to find out what population censuses have been taken in the world since (say) 1970—and to get a complete and accurate list within minutes.

A developing country researcher studying poverty should be able to learn, again with a single, simple query, how many censuses, already taken or now underway, include a question on income—and to learn by simple follow up queries how the income question was asked, what non-response rates were, how the data were edited, what tabulations were published, in print or on line, and where the data can be obtained, if not available on line.

A demographer working in the Population Division of the United Nations Secretariat in New York needs to update age-sex distributions from population censuses for the 2000 Revision of the biennially revised *World Population Prospects* reports. Because the data files follow standard naming conventions, the same conventions used by national statistical office web servers all over the world, a simple world wide web “robot” computer program accesses the national statistical office websites (nearly 200 of them), checks to see if a more recent census age-sex distribution is available and, if so, downloads the file. (*C’est moi.*)

This is what we are aiming for: a radical improvement in the ease and speed of access to all kinds of census information. To achieve this it is necessary to automate the work, that is, to have it

carried out by computer programs rather than manually. And to achieve *this* it is necessary to adopt some sort of minimal standard for the organization and naming of census information.

That is what this paper is about. And before you say that it is impossible, consider the precedent of the world wide web: a non-proprietary protocol, implemented with freely available software, developed by a relative handful of people, that has in a mere ten years (the length of an intercensal interval ...) blanketed the world.

### **What's in a name?**

*Information!* Names may go beyond naming to provide information about the things they name. Street names, for example, may be purely notional, so that “Elm Street” tells us nothing about Elm Street except its name, but they may also encode information about the location of streets in relation to other streets. This is the case, for example, for the numbered streets and avenues of New York City.

The naming system suggested here draws on the model of the internet domain name system (DNS). The DNS provides a unique name to every one of the many tens of millions of computers all over the world connected to the internet, but it also provides important high level information about the locations of the named computers in the Internet.

The computer named **stats.gov.cn**, for example, tells us that the computer was registered in China (**cn**), belongs (probably) to the government (**gov**) and (presumably) has something to do with statistics (**stats**).

Now let's say a bit about each of those technical words in the title of the paper.

### **Universal**

The idea is to develop a system for assigning a unique name to *every* significant population census information resource, for *every* census ever taken or underway, in *every* country in the world. “Information” refers to all pertinent data and metadata—census schedules, codebooks, data files, publications, definition of the population enumerated, contact information for the census organization. (It should be recognized that there is no simple, unambiguous line between “data” and “metadata”.)

The significance of universality is that any national statistical office will be *able* to use the system and will, if they choose to adopt it, make it easier for themselves and their users to know what census information exists and to access this information.

### **Object-oriented**

Those of you who follow software engineering know what a loaded term “object-oriented” is, and also that it is generally regarded as the way of the future. This would be an unsuitable forum in which to discuss object oriented design, programming and data base concepts, but one particular practical significance of object-orientation in this context is simply that it provides a means of insuring that data can't get separated from it's metadata. Which is a good thing because data isn't very (or perhaps at all) useful when separated from it's metadata.

## Hierarchical

The concept of hierarchy is dear to the heart of census takers in the form of the geopolitical units that are used to partition the censused area into progressively smaller pieces, ending with census enumeration districts.

This is not the only hierarchy in census data, however, or even the most important one. Essentially all census data is attached to a particular country and to year in which the census was taken, so it does not take great imagination to invent a suitable name for particular censuses, such as

**china.census.1990,**

to name the 1990 population census of China.

We think of all the information for a particular census as being a “census data object” named according to the scheme **country.census.year**. A census data object is large and complex, consisting of a large number of component objects. One very simple, “atomic” object is the reference time for the census, which might be named, e.g., **us.census.2000.time**. Data for persons and households are themselves quite complicated data objects, but they might be simply named as, respectively,

**us.census.2000.persons**

and

**us.census.2000.households.**

## Extensible

No scheme that requires anticipating everything in advance will work, because we can’t anticipate everything in advance. A usable scheme must incorporate procedures for being readily extended to accommodate new needs. This is what we mean by “extensible.”

If a census collects information on “communities”, for example, (whatever exactly this may mean) it must be possible to define a census data object to contain community data. This object might be named, e.g., **indonesia.census.1990.communities**. Its existence could be declared in an object named **indonesia.census.1990.aggregates**, and it could be defined in an object named **indonesia.census.1990.communities.definition**.

## Conveniences

When working on (say) the 2000 census of China **china.census.2000** may be implicit, so we don’t have to bother with writing it every time. Human contexts may be ambiguous, but computing contexts need not be so, as exemplified by “present working directory” and “environment variables”.

Clearly we will want to allow *aliases*, so that we can write, for example, **us** rather than **united-states-of-america** and have the system know what we mean.

“Wild cards” will be useful also, allowing us to write, for example, `china.census.years` to refer to all censuses taken in China, or `countries.census.1990` for all censuses taken in 1990.

### High level “person” data set objects

At the highest level, person data from a census involve three elements: `persons.aggregate` will be an overview document describing the population enumerated, with references to more detailed information; `persons.items` will be a (relational data base format) table listing all person information items; and `persons.values` will be a row-column, person-item array indicating codes for each information item for each person. For each item in `persons.items` there will be a (relational data base format) table showing the codes and values for this item, e.g., `persons.age`, `persons.sex`, `persons.rhh` (relation to head of household).

### Tabulation objects

Tabulation objects for persons may be named, e.g., `persons.tabulation.age-sex`. The *domain* of the tabulation, the set of persons cross-classified, is implicit in the domains of the items that are the dimensions of the variables. It need not, therefore, be indicated explicitly. The domain of the table will always be the intersection of the domains of the items that define it.

### Think globally, act locally

Realizing the potential of information technology to facilitate easy, rapid, automated access to population census information requires cooperation between individuals and organizations throughout the world. The world wide web illustrates the best hope for realizing this cooperation: recommended protocols and freely available software tools for implementation, voluntarily adopted because adoption serves a wide range of particular interests.

But we need not wait for “world wide” implementation to make use of and further develop the idea of a uniform naming system for population census data resources. Such a system can be a useful tool for organizing population census work in any national statistical office without being used in any other office. It can put to good use in any part of an organization, or even by a single individual seeking to organize his or her files.

This was in fact the way in which this “modest proposal” began. Collecting basic population data for population estimates and projections, I needed to assign file names to files containing various kinds of data. What names to use? Ad hoc naming schemes make things difficult for others (and often oneself, in the future), and most computer operating systems now in use allow long file names (how long is more of an issue than one would like, but this is not a detail to pursue here). Why not assign names that describe, with the help of some simple rules, the contents of the files?

Having done so, I realized that this makes it possible to automate various operations that would otherwise require “manual labor” at the computer keyboard. I wrote, for example, a simple program that searched for all available census age-sex distributions for a given country and plotted the corresponding population totals against time.

## Conclusion

To conclude, let me point out that this really is a *modest* proposal, because it refers only to *population census* data. The same ideas may be applied with equal validity (it is of course for you to judge what this validity may be) to population survey data, to vital registration data, or to all demographic data. Or, indeed, to *all* (or at least a very wide class of) data. *That* would be an *immodest* proposal. Thank you for your attention.

## References

*Weaving the Web*, Tim Berners-Lee with Mark Fischetti, HarperSanFrancisco, 1999. An illuminating account of the development of the world wide web, by the person who started it all. (No, that's not a typo, that's really how the publisher's name is shown in the book.)

*Futurize Your Enterprise Now*, David Siegel, John Wiley & Sons, Inc., New York, 1999. A futuristic look at how information technology might shape society. You may not be as enthusiastic about the prospects as the author is, but the book is valuable for the breadth and detail of the imagined possibilities.

*Database Nation*, Simson Garfinkel, O'Reilly, Beijing, 2000. Tangential to this discussion in some ways (but not in others, privacy being a major issue in census taking), but Siegel comes across as so blithe on the issue of privacy that an antidote may be required. This book provides one. Ironically, the population census is blamed for starting it all. Why? Because the data processing demands of the United States census stimulated development of the Hollerith ("punch") card. Remember those?