Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

Merging data from a collection of Catalogues

Patricio F. Ortiz

Department of Physics & Astronomy, University of Leicester

Abstract.
Computer-assisted catalogue-merging utilities are the natural extension to manual methods used by astronomers to merge information from different catalogues in the pre-VO era. Recognizing which columns represent the same physical quantity is important not only to combine two or more tables, but it is of prime importance when the aim is to recognize sources with similar physical characteristics in an N-dimensional parameter space, ie, true data-mining. AstroGrid's Data Federation Research Group studied the problem of merging columns from two or more sources (VOTables) based on their meta-data in order to merge the results of several cone-searches or from cross-match of catalogues. The results presented here are relevant in future usage within the VO.

1. Introduction

The problem of merging information from diverse sources is not a new one, astronomers who have prepared compilation catalogues are familiar with the issues, and their way to solve the problem should shed a light on the methods we devise for the future, when the data avalanche arrives.

Today's data-centres may hold several thousands catalogues, some with a few columns, others with hundreds of columns; some with a few data points, some with zillions of data points. The astronomical literature is populated with tables which tend to cluster according to their content, therefore, there will be cases where the intersection between a set of catalogues is empty, for example:

RA	Dec	Vmag	B-V
hms	dms	mag	mag

ObjectID	SpType	XFlux	z
		Jy	

The more interesting cases are those where the intersection is not empty. The question of whether two quantities are equivalent or not is not a simple one to answer though. Let us consider the following situation:

RA	Dec	Vmag	Flux_X	z
hms	dms	mag	mW/m2	kpc

RAB1950	DecB1950	XFlux	z
deg	deg	Jy	

this example is the worst case scenario. RA and Dec can be identified as similar quantities, but they are represented in different units, and worse, they may represent values for different equinoxes. The case of **Flux_X** and **XFlux** is

worse, we don't have enough information about what they represent and the units are different. Are their units scalable? The case of **z** is no better. The first one is a distance, but what does the second one represent? Redshift? A relative distance?

The solution in the case of manual merging is to resource to additional information, table captions or the paper itself. Currently, catalogues are detached from some of these pieces of information, even if a standardized description (ReadMe) file is attached to them.

Modern day merging requires an additional piece of meta-data: the UCDs or **Unified Content Descriptors** (Ortiz et al. 1999, Ortiz 2000). The following scenario (where a descriptor is added to each column) is better:

RA	Dec	Vmag	Flux_X	z
hms	dms	mag	mW/m2	kpc
pos.eq.ra,main	pos.eq.dec,main	phot.jhn.v	phot.flux.x	phys.distance

RAB1950.0	DecB1950.0	XFlux	z
deg	deg	Jy	
pos.eq.ra,main	pos.eq.dec,main	phot.flux.x	redshift.hc

Better, but not perfect. However, it is clear that by using UCDs it is possible to recognize the quantities which are the same much more easily than before. **z** does represent different quantities; X fluxes share the same UCD but have different units. The case of RA and Dec is still ambiguous.

Note that UCDs act as another piece of meta-information attached to each column in a catalogue telling us the most likely nature of this quantity.

Reality shows us that in many cases authors choose to represent quantities in linear scale while others use logarithmic scale. Currently there is nothing contained in a VOTable to tell us in which scale a quantity is measured: we may have to rely on the column explanation, or the name (logMass as opposed to Mass), or in **Vizier** (`http://vizier.u-strasbg.fr/cgi-bin/VizieR`) the use of square brackets around the units. There is no standard today, yet **data scale needs to be part of a column meta-data**

The ambiguities do not end there: some quantities are represented with a zero point subtracted; particularly worrisome is the case of the dates, in which it is not rare to see "date" measured from 1950, 1975, 2000, the author's favourite equinox, or the date of the first observation in a run. We suggest that a **zero point** should be present in the meta-data (or data-model).

2. Experimental setup

To investigate the full-scale problem based on facts and large numbers, we decided to download the meta-data from one of the largest catalogue collections available today: **Vizier**, from the Centre de Données de Strasbourg (CDS). Vizier holds about 10000 tables with nearly 125000 columns, most of them with attached UCDs. The lessons learned should be applied to fully automatic and computer-assisted methods to merge information.

In our first approach we selected a set of catalogues based on their meta-data (at the table or column level), to obtain the list of used UCDs. We arbitrarily

picked UCDs from the list and produced the list of the columns associated with them; the units used were of particular interest. This showed that in some cases, quantities tagged with the same UCD were measured in different units, mixed linear and log scale or had different zero points.

We found out that the most reliable solution to determine if two units were equivalent was to use dimensional analysis. Each unit can be represented as a combination of SI components (kg, m, s, etc), therefore one has the situation:

$$unit_a = factor_a \times SIeq_a$$
$$unit_b = factor_b \times SIeq_b$$

if $SIeq_a = SIeq_b$, then the two units are equivalent and there is a conversion factor between both of them ($factor_a/factor_b$), if not, the quantities likely represent something different. We used the CDS program **units**, by François Ochsenbein (`http://cdsweb.u-strasbg.fr/viz-bin/Unit`) to obtain the SI conversion factors and equivalences.

3. Results

Our experiment with the meta-data proved that: **a) UCDs** in combination with **units** is highly reliable to determine if two columns represent the same quantity, **b)** federating/matching is possible even if the units do not coincide, **c)** dimensional analysis should be used to determine the equivalence of two quantities when the units do not coincide.

It is also clear that: there are no "politically correct" units to measure a quantity, users should become involved in the process of combining information from various sources and eventually, all meta-data should be made available to them in future tools to develop (column names, units, explanations, UCDs, and others to come). Automatic merging is possible, but there are many cases in which only humans should decide how and what to merge.

In order to provide a much more complete picture to those who want to merge information some key pieces of meta-data are still missing, and we should seriously consider to incorporate them to **registries** as their presence will make the situation clearer for users and developers. The most important ones are: type of representation (**linear/log scale**), presence of a **zero point**, **scale factors** (some services absorb this in the representation of the units), **equinox/epoch** of equatorial coordinates (some provide this within resulting VOTables and it can be combined with UCDs to determine equivalence of equatorial coordinates), and finally, **minimum and maximum** value of any given column (already proposed in IVOA documents).

Some other important issues we discovered include: **i)** a significant number of catalogues contain object identifiers but no coordinates. These catalogues will be invisible to positional search engines despite how important many of them are. **ii)** UCD assignation method should be more rigorous as the task is difficult when column explanations are cryptic or ambiguous. **iii)** The units used to measure proper motion in RA are very important to compute precession, yet, the $cos(\delta)$ factor is present in some and absent in others cases without clear indication in the description files.

4. Impact on "Urbe et Orbi" queries and future work

A side effect of our experience is that a number of issues concerning generalized queries to VO services came to light. Let us consider the following request:
Select all entries where UCD(REDSHIFT) is between 0.5 and 1.7 and UCD(RA) > 10 Retrieve UCD(RA) UCD(DEC) UCD(REDSHIFT) UCD(BRIGHTNESS)
this request makes sense for an astronomer, but this type of query needs a lot of translation and interpretation for an automatic system to answer it. The critical aspects dealing with this query have to do with picking up the catalogues which contain the relevant information and satisfy the user constraints.
• **UCD(REDSHIFT) is between 0.5 and 1.7** is relatively easy to interpret, REDSHIFT probably means heliocentric redshift so we can look for columns which represent that quantity and select the points where the value is between 0.5 and 1.7. • **UCD(RA) > 10** is ambiguous. RA in which equinox? Should we assume J2000? What is 10? 10 hours, 10 radians or 10 degrees? What value should be passed to the DBMS? If the number 10 is passed bare and it is interpreted in "local units", the results will be mixed between degrees and hours. **Units should then be attached to the query**. We should say **UCD(RA) > 10 hours**. • Assuming a set of "standard" units is dangerous, as we will never agree how to measure things (metric vs imperial is a good example). Users should express quantities in whatever units they please, as long as there is a layer which provides translation. • Registries (or other component of VO) should then take care of converting these units to those which each catalogue is measured in before passing the query to the DBMS **or** services should accept queries in which each quantity is represented by a number and its units internally translating the query. • Column names are unique in a catalogue, UCDs are not. How do we select a column among several which have the same UCD?

Several groups are currently working in the area of data-merging with different emphasis and perspectives, (Page, 2004), (Adorf et al., 2004), and other groups around the world. What seems to be very important is to have some working prototypes which will allow us, together with users, to identify the most relevant issues. The problem of merging data is not a simple one to solve and there is no unique solution, but it is absolutely relevant to fully exploit the possibilities that the VO opens, as it also applies to studies in the time domain.

Acknowledgments AstroGrid is funded by the Particle Physics and Astronomy Council of the United Kingdom.

References

Adorf, H., Lemson, G., Voges, W., Enke, H., & Steinmetz, M. 2004, this volume281

Ortiz, P. Ochsenbein, F., Wicenec, A. & Albrecht, M. 1999, in ASP Conf. Ser., Vol. 172, ADASS VIII, ed. D. M. Mehringer, R. L. Plante, & D. A. Roberts (San Francisco: ASP)

Ortiz, P. F., 2000, in Computer Physics Communications, Vol 127, p. 188.

Page, C. 2004, this volume189