How Accessible are Extragalactic and Radio Data from Journal Tables?

Heinz Andernach

Dept. de Astronomía, Univ. Guanajuato, Mexico

XXVI IAU GA, SPS 6, 22 Aug '06

Parallel to the collection at established data centers, and partly for my own research, I collected since 1989 the tabular data from over 2100 articles, half of these concerned with radio sources, the other half with extragalactic objects in general. Optical character recognition (OCR) was used to recover tables from ~470 papers. Tables from only about half of the 2100 articles are available from either the CDS or CATS (cats.sao.ru) catalog collections, and a similar coverage is estimated for NED. Both, object databases (NED, SIMBAD, LEDA) as well as catalog browsers (VizieR, CATS) complement each other. More human resources at the data centers and better collaboration between authors, referees, editor/publisher and data centers are required to improve data coverage and accessibility. The VO project should move a moderate amount of emphasis from interfaces and visualization to data content.

A One-Man’s Data Center . . . Why?

Based initially on a lack of radio source data in NED and SIMBAD, since 1989 I collect electronic tables of radio sources and/or extragalactic objects that were not readily available from data centres. My collection is currently growing at ~200 cats/yr (1 cat = 1 article with at least 1 table):

The Catalog Collection at CDS

The CDS collection currently grows at ~500 cats/yr (solid line, Fig. 2). As some are unsuitable for cone searches (e.g. for lack of coordinates), not all, but ~87% are in VizieR (dot-dashed line). Remarkably, for papers published ~1982–92, my collection (dashed line) contains almost as many cats as CDS, probably due to my OCR’ing of tables from that period, but CDS’s and my collection overlap only to a small extent! CDS catalogs listed as “in preparation” (dotted line) have a median age since publication of 5 years, twice the time I derived in mid-2003 (cf. www.inaoep.mx/~survey03/heinz.html): a backlog piling up!

Pre-1960 catalogs at CDS are mainly stellar.

Catalog “Biometrics” and Data Center Coverage

The size distribution of both radio- (Fig. 3) and non-radio (Fig. 4) catalogs follows a power law (known as Zipf or Lotka laws in biometrics) with almost equal slopes, ~0.68 and ~0.65, respectively. The turnover for small sizes < ~100 records indicates incompleteness of the catalog collections. While most of the large radio catalogs (> 1000 records) are included in either CATS or the CDS collection (note that these collections differ)...

Fig. 3: Size distribution of 1218 radio catalogs

... this is different for non-radio catalogs in my collection:

Fig. 4: Size distribution of non-radio catalogs

Only 38% are available at CDS, growing to 50% for >100 records, and 58% for >1000 records. These percentages increase only slightly when a fraction of items in my collection are discarded which may not be suitable for archiving at CDS, e.g. for lack of adequate documentation.