HERITAGE ON TAPE (Inaugural address)

Wulff D. Heintz
Dept. of Astronomy, Swarthmore College

For your friendly welcome, Dr.Florsch and Dr.Curien, please accept our sincere thanks. The astronomers specialising in data documentation, and their Commission in the IAU, are very grateful indeed to be able to meet here again after our first conference in 1976. It is also an honor that this Colloquium with participation from 15 countries is leading off a series of events celebrating the centennial of the Strasbourg Observatory. The addition of astronomy to your academic community has been one of continuing success, characterised by a long list of important publications, and of well-known names associated with the place through the decades. Thus you possess an active center of research and documentation in astronomy, and this in a charming, hospitable city of extraordinary beauty. Certainly it is to be ascribed to the attraction of the place and to the renown of our hosts as well as to the important subject of the Colloquium, that your invitation has drawn so large a response.

Between the generation and the use of our products lie publishing, abstracting and tagging, collection, intermediary or selective evaluation, repackaging, and dissemination or marketing. This quotation does not come from an economics textbook on merchandise, but from a CODATA publication. Our research experience has acquainted us not only with the production and application of data, the ends of the chain, but also with the information brokerage steps in between, which simply are inevitable once a society and its supply/demand transaction rate reach a certain size. This applies to supermarket sales as well as to star positions. Astronomy with its long history, and with the longevity of gathered data has early become aware of the problem, and it was already in 1921, in preparation for the first IAU assembly, that M.Baillaud of the Paris Observatory was invited to form a Commission which - first of all at that time - was to concern itself with bibliography and abstracting. Thus we have reached, along with the Strasbourg Observatory centennial, the 60th anniversary of Commission 5. We have witnessed the strong additional emphasis on data documentation, storage, and processing in the last 15 years, when computer systems had become available to do large-scale jobs in that area, and we have met here five years ago to develop

C. Jaschek and W. Heintz (eds.), Automated Data Retrieval in Astronomy, xvii–xx.

ideas on what can be, and ought to be, done by way of coordination. The technical development since has been so breathtakingly fast, the scientific personnel involved also increasing, that another conference seemed desirable, and its initiator, C.Jaschek, needed little persuasion to get it approved.

Much has happened in the meantime, as can be seen from the data center bulletins and other reports. Machine-readable holdings have substantially increased. The spreading of computer terminals linking with databanks and of networks between centers puts a much larger clientele into reach of the data which they need, and can now obtain wholesale. B.Hauck reported on this contribution from astronomy at the last CODATA meeting. The 75-page listing of existing facilities by Jaschek at our preceding conference would now certainly be thicker. Important progress has been made toward an object nomenclature and a controlled vocabulary for the purpose of data tagging with good reliability of completeness. We have at least the promise of various IAU Commissions to try and promote such standards, and it remains to convince also the publishers of primary literature - most of which is not in astronomical hands - where sometimes vague and inconsistent publishing policies can be sources of mistakes. Toward this goal the efforts toward standards for presentation of data in primary journals along CODATA guidelines should also prove helpful.

Scientists in Physics, Chemistry, and Mineralogy, for instance, require all kinds of properties of a million compounds and minerals to be contained in their databanks. Most of such data are time-invariant and reproducible, making revision and augmentation - along with an elaborate multi-purpose access - the main tasks for the compiler. On the other hand, geosciences have been flooded - particularly through satellites - with huge amounts of instantaneous data. One weather satellite yields 10^7 binary digits per transmission, or something like 10^{12} bits of output per year, and the synoptic readings, on which one weather analysis is based, contain 10^6 to 10^7 bits with a required compiling/processing time under 1 hour through the World Meteorological Network. But many of these data are of transient significance only, and need be stored merely in condensed form for long-term reference.

Astronomy has an in-between position, facing various requirements for its various needs. Large data material is (or may be assumed to be) time-dependent and not repeatable, and the interval between acqusition and perusal frequently is long. We are now in a position that large enterprises like the Palomar Survey or the 21-cm mapping could soon be repeated and automatically compared with first-epoch tapes for variations. Given also the interdisciplinary nature of much of the research, particularly in Astro- and Space Physics, we should provide for a data tagging system with as much compatibility as we can afford without sacrificing those classification features which we - although a smaller community than Physics - deem indispensable for our research approach. And finally, since Astronomy has always been a struggle with noise levels, caveats on the degree of reliability of data need special attention. In older observations the lower precision may outweigh the longer timespan.

A related issue recently caught my attention, namely the guidelines for abstracts. They could be improved in some ways: homogeneity, emphasis on data flags and on inexpensive, limited-character reproduction by abstracting services, and enforcement by the primary periodicals.

To those who instruct students and advise younger colleagues, a suggestion may be addressed: Make them aware of the means - bibliographic and data storage - that are at their disposal. Part of the scientific community does have a traditionalist attitude toward new mechanisms in information, the shelves of printed books and catalogues are deemed irreplaceable, and there may be some doubt involved as to the quality and completeness of the material offered via terminal. Some commercial reference systems in USA are indeed still deficient in completeness and in cross-references. Yet the next generation of scientists - having had the computers already in the cradles - will have to depend on the terminal along with the microfiche reader, and these tools will enable them to work with better efficiency and yet with the same thoroughness, in comparison with the potential of earlier gnerations. The current literature on data retrieval is partly management-oriented, full of codes, acronyms and other documentese language, and not easy to read. However, we have user-oriented communications from some data centers, and retrieval references could gradually emerge as vital parts of any advanced textbook, much as bibliographies for recommended reading have long been already.

When preparation of this conference began a year and a half ago, suggestions for the agenda piled up so rapidly that we had to make a selection from the shopping list, lest we would have had to impose upon our hosts for another week. One of the items dropped from the agenda was "Financial Trends", on which - beyond a certainly very stringent awareness - not much can be done by way of coordination. I believe that something should be noted, and considered on the long run: It has been estimated that, even in highly industrialised countries, a few tenths of 1% of public Research and Development funds at best go into data processing and retrieval. At this level we cannot expect to do justice to the material to be processed, and the wealth of information is answered by a poverty of attention. It seems that 2% or 3% of R/D expenditures would be a more adequate figure, and still 10 times less than what science management requires. Compare this with the gains to be achieved if at least part of the time and costs for inadvertent and redundant duplication of data could be avoided through expedient access to pre-existing material. Needless to say, a larger retrieval clientele will also reduce the unit costs of search.

As to which structure the combination of databases will take, some thoughts offered by G.Wilkins at the close of the previous conference can be drawn upon. The data centers at the heart of the system are primarily charged with the responsibility for data of the publishable kind, that is, published at least under the standards of golden times before the invention of page charges. In addition there will be datafiles at research institutions, not present in the centers except by reference. These will have the least access demand, yet they will contain the oc-

casionally needed backup information. It would usually still be uneconomical, for instance, to put photographic archives on tape. (This situation is comparable to the difference between abstracts, taped and disseminated by abstracting services, and full-length papers which require a subscription or a copying service. Some sciences outside Astronomy even got into the unfortunate position that full papers cannot be printed any more but have to be retrieved from a depository.) Thirdly, we have the derivative, selective, or critically evaluated data from the hands of the specialists of respective subject areas, a kind of synoptic literature. These files will receive most access requests, and from a most diversified and nonspecialist clientele. They should become fully banked, probably also accompanied by explanations of the compilation or by caveats of use. It is at this point that advice from designated specialists, Commissions, and Working Groups must be relied on most heavily, in the modification of accessible data, and also in the termination of access, should for instance a classification scheme or a set of standard objects be deleted in order not to cause future confusion. Where this cooperative effort has still been deficient (perhaps owing to inertia or to the desire to stay out of conflicts), we may proceed on the optimistic assumption that, as the importance of data services grows, so will the interest of the scientists to have the most reliable, complete, and error-free data material on record.

Newton's First Law states that everything in the world is determined by inertia, unless an effort toward change is made. Of course this applies only to the world of physical science and not to society. Yet we are far enough into the computer age that society does realise the advantages of modern information chanelling. The problems we are to negotiate are of profound influence on the structure of scientific communication. They relate to the spectrum of different customer demands, also to interscientific and international compatibility, standardisation, efficiency of management and guidelines, and far-sighted planning. Since problems, however, are not getting solved by reiterating their complexity, let me conclude, again appreciating your attendance here, and let us get to business.