LISA V - Library and Information Services in Astronomy V
Common Challenges, Uncommon Solutions
Cambridge, Ma, 2006
draft of the paper submitted for the proceedings
ASP Conference Series, Vol. 377, proceedings of the conference held 18-21 June 2006 in Cambridge, Massachusetts, USA. Edited by Sandra Ricketts, Christina Birdie, and Eva Isaksson., p.47

Observations and publications in the VO: is the VO only for Big Science?

A. Holl
Konkoly Observatory, Budapest, Hungary

Abstract

The Virtual Observatory (VO) got started by Big Science projects. We can ask: is the VO an exclusive tool to handle big catalogs from large surveys? We think the VO should contain information from journals, other publications, and non-survey, "Little Science" observations too. Journals, observatory publications and repositories could be the vehicle to carry non-survey observations to the VO.

Is the VO applicable for Big Science data only?

Big Science is where the VO was born. Handling Big Science data is relatively easy for the VO: surveys yield large quantities of uniform data. Big Science projects usually have large budgets and it is easy to incorporate meta-data into the process.

But what about "Little Science"? Would not it be a loss for the astronomical community if data from small projects did not get into the VO? We think that the small could be important. An argument for building the VO is that the secondary use of data means greater efficiency in terms of scientific results for the same investment. If efficiency is important, we can not afford to lose even the smallest bit of good quality research data. The loss of visibility resulting from possible exclusion from the VO would be grossly unfair too. C. Sterken deals with the questions of Big Science and Little Science Sterken (2000), and quotes Weinberg(1961):

"We must make Big Science flourish without, at the same time, allowing it to trample Little Science - that is, we must nurture small-scale excellence as carefully as we lavish gifts on large-scale spectaculars."

But there are other, maybe more important arguments in favor of including small project data in the VO data pool. These data are mostly good quality (of course, regarding only data which have gone through peer review or other quality control already): they are hand-made (hand-reduced) by experts. All the small observatories, and small astronomical projects together might add up to a large survey.
With exponentially growing data volume it is true that new big project

data quickly outgrow the existing ones - but the time dimension matters! We cannot repeat lost observations even with the most modern instruments. Data resulting from small projects could be important in filling the "information gaps".

What vehicles could carry small project data to the VO?

We can see four possibilities: journals, observatory publications, repositories and annotation of databases.

- Journals

Publications and the data they are based on must be both accessible to the reader and the referee. Of course, it could be accomplished by referring the data in a public archive. Standards are suggested for this by Accomazzi (2004). In turn, databases should contain links to the publication based on a given dataset too. Data associated with papers published in A\&A are stored at CDS - VO inclusion is solved in this case.

Journals themselves could store observational data. This is the case for the small journal "Information Bulletin on Variable Stars". IBVS already uses some VO techniques Holl (2004b), and we intend to add more such features.

Journals, of course, contain data besides observational data files. How could these data get into the VO? Presently, information in the journal articles will get into Simbad or other databases manually. We think with some modification of the way astronomy papers are coded, it will be possible to harvest the data in published articles automatically.

Fig. 1. Information published in IBVS is entered to databases and bibliographies - a process which could be automated.

We think that journal articles should be readable not only for humans, but machines as well. At this point, as a thought experiment, we can consider changing the data format of the scientific papers from the present LaTeX and PostScript (for figures) to XML. Transforming the ideas of Donald Knuth for literate programming (Knuth 1984), we can envision a scientific paper, where text, data and figures are coded in a way which, if processed for the human reader, would result in a nicely typeset paper. On the other hand, all the tables and figures would be machine-readable. The difficulty is, of course, that we would need proper tools not only to format these papers, but to prepare the manuscript as well. The advantage would be that information from such papers could be channeled to databases more easily, and there would be no more need for retrieving data from a published figure using a ruler. A further advantage could be, as Knuth described, the improved quality of the papers, because the authors would find it necessary to present their data more carefully.

Without "literate scientific data presentation", is there anything journals can do to help their data being included in the VO? We think so.

Firstly, they should use as many tables as possible. It is easier to harvest information from tables than from text. Format the tables for machines or provide a machine-readable version as well, and make allowances for VOTable conversion, e.g. including UCDs, maybe in a way that they do not appear in print. A few selected tables at IBVS are already available in VOTable form as well.

Secondly, journals can encourage authors to submit data they plot on figures in tabular form as well (electronic journals can store lots of supplementary material).

Furthermore, it would be important to have all tables, figures, data files published in a journal accessible and addressable. At IBVS, all published figures can be re-used, included in other services, by using a unique identifier composed of the issue and figure numbers, e.g.

  http://www.konkoly.hu/cgi-bin/IBVSfigure?5656-f1&Format=compact

For this, every figure, table and data file needs meta-information including some special keywords. We intend to add database-like functionality to the electronic IBVS Holl (2004a). The content of a journal article is static after publication (except for the possible errata, the existence of which should be indicated). On the other hand, meta-data should be dynamic. At IBVS we have taken efforts to identify "anonymous" objects in old papers, and update the meta-data files with modern identifiers.

Finally, good markup could also expedite metadata harvesting. An example is the \SIMBADobj LaTeX macro at IBVS.

- Observatory publications and archives

Observatory publication series should be continued (Holl & Vargha 2003). We argue that observatory publications should go hand-in-hand with observatory archives. In the past, huge datasets and catalogs were published in observatory publications. This can only be done electronically now. But electronically published catalogs - as librarians know very well - lack something paper catalogs have. Electronic catalog handling systems, like Vizier at CDS, often discard old versions superseded by new ones. And authors citing electronic catalogs or tables often do not use unique computer identifiers, but publication dates instead. We find it useful to keep the old way of publication year, series, volume for reference, and have a handle for every electronic catalog on the library shelves too (which could be entirely virtual). Electronic catalogs benefit from having a publication counterpart, but observatory publications, on the other hand, should become electronic (not excluding, even recommending a printed version).

Creating and maintaining meta-information is an essential task for observatory publications and data archives alike. Librarians might have an active role in this - this is where they have considerable expertise. As for the contents of the websites, archiving is a relatively new concern. We have put electronic databases and catalogs on-line, and by now we have suffered our first losses:

databases and websites which ceased to be accessible, or not maintained any longer. We need librarians to keep essential information in these. Publishing some descriptions of these electronic resources in observatory publications might be a way of archiving bits of these resources.

There is an important lesson we can learn from the Open Archives / Open Access community. Observatories now can get shrink-wrapped software, to maintain publications electronically, like EPrints (University of Southampton) or DSpace (MIT). OAI technologies for metadata harvesting could be used.

We would need similar software for data archives. Small observatories cannot afford to build their VO compatible archives - there is need for easy-to-install and -maintain ready-made software, into which CCD direct frames or spectra could be uploaded, and which provide, besides the basic database functionality, communication with the VO with standard protocols like ConeSearch and others. A possible candidate software for this task is SAADA developed at Strasbourg Observatory (Nguyen 2004). EPrints and DSpace could also be used for data repositories - its applicability in astronomy should be investigated.

- Data repositories

Is there a need for repositories which would accept small datasets from professionals, or others open for amateurs? We think the answer is yes. NCSA started a kind of repository years ago: the NCSA Astronomy Digital Image Library (Plante 1996)
It was created to store fully processed images, and is populated with a relatively small amount of nice pictures mainly from the field of radio astronomy. The repositories we propose would store not only nice images for outreach, but could archive and open up for secondary usage of reduced CCD frames or spectra from any professional observatory. The advances in technology would make it feasible to store all amateur CCD images submitted in a proper format and being equipped with the necessary meta-information. The application of EPrints or DSpace could be considered here as well.

The crucial question here is quality. But we believe that for simple data - like direct CCD frames - quality control could be built in the process.

- Annotated VO archives

In variable star astronomy there are large datasets of time series photometry available for the public, like NSVS which is available through SkyDOT (Starr 2003) or ASAS (Pojmanski 2003).
These contain photometry which was processed automatically, searching for new variable stars. Lightcurve parameters assigned by pipeline processing are far from being accurate, and the variability of great many objects remain undetected by the software. Amateurs and professionals mine these open databases, finding further variables and refining (or completely changing) the classification and parameters of others. Results of such investigations appeared frequently in IBVS. The caveats can be clearly seen: while some researchers (amateurs and professionals) with proper knowledge and care produce good science

from such data mining, others derive questionable or poor results.

Results of the utilization of publicly available data could be fed back to the VO through the journals, like IBVS. Another interesting possibility is direct feedback to the open database in the fashion of the popular Wiki. The important question is, again the quality control - such feedback should only be allowed for selected researchers probably.

Acknowledgements: The author is grateful to "Friends of LISA V" for their help.

References

Accomazzi, A., Eichhorn, G., 2004, ASPC, 314, 181

Holl, A., 2004, ASPC, 314, 229

Holl, A., 2004, AN, 325, 610

Holl, A., Vargha, M., 2003, in: Library and Information Services in Astronomy, eds.: Corbin, B., Bryson, E., Wolf, M., U.S. Naval Obs., p. 109

Knuth, D., 1984, The Computer Journal, 27, 97

Nguyen, N.H., Michel, L., Motch, C., 2004, ASPC, 314, 121

Plante, R.L., Crutcher, R.M., Sharpe, R.K., 1996, ASPC, 101, 581

Pojmanski, G. 2003, Acta Astronomica, 53, 341

Starr, D., Wozniak, P., Vestrand, W. T. 2003, in ASP Conf. Ser. Vol. 295, Astronomical Data Analysis Software and Systems XII, ed.
H. E. Payne, R. I. Jedrzejewski & R. N. Hook (San Francisco: ASP), 85

Sterken, C., 2000, IBVS, No. 5000

Weinberg, A.M., 1961, Science, 143, 161