Blue skies and clouds, archives of the future

William O'Mullane

Issue 01.0 2011-04-06

1 Introduction

Hipparcos had a huge impact on astronomy and even further afield 2009. Gaia should have an even bigger impact. But the final Gaia catalogue will not hit the net until 2021 or so - how might this look? Some current ideas are bubbling and some general points are discussed. Many thanks to the various GAP,DPACE and GST members who provided comments.

2 Shovels and Bulldozers - advanced statistics

Many astronomers today research small portions of the sky and most archives are well capable of answering queries for such matters. The spatial nature of astronomy means that this type of query will always be required. But we are in the era of mega surveys producing huge amounts of data across large sky areas (or the entire sky like Planck and Gaia). With large homogeneous data sets, statistics become important, now an astronomer interested in a particular object may want to search for *similar* objects in a bunch of surveys. This is hard and needs to be automated - the virtual observatory does not yet provide anything capable of doing this. If we take as an analogy that most current astronomers are digging with shovels, some are realizing they need a bulldozer but no one has built it yet. We need to work on advanced statistical techniques and making them available on our archives. Precomputing some statistics and allowing the user to query those to reduce the size of the dataset of interest is a start. Ultimately for any part or all of the data a user should be able to use cluster analysis, pattern recognition and N-point correlation function calculations.

In addition we should be going toward near natural language as the interface - for example, we could consider asking Wolfram research to make the archive with a Wolfram Alpha like interface. One could then imagine an alpha type interaction like ”plot magnitude against and colour for all stars with parallax between 1 and 1.5”. Intersystems are going a similar way with iKnow[1] and DeepSee. With iKnow one could consider loading all available journal articles and making them available as part of an archive search - the technology is quite interesting, it is not simply a google like search it is far more accurate taking account of linking concepts. In any case it is worth looking outside astronomy, Gapminder `http://www.gapminder.org/` for example has an intuitive plotting interface.

Could we consider having a ”Scientific Assistant” (Scientific American January 2011, p. 59) with artificial intelligence built in ?

[1] `http://www.iknow.be`

3 Three Dimensions - at least

Astronomy data is obviously multi dimensional, the spatial nature makes many think of three dimensions and a sphere for positional information. The dimensionality is in fact much higher - one may think of any catalogue as a multi-dimensional space of all attributes in the the catalogue where each attribute represents a dimension. If we could sort in this multispace, similar objects would appear *close* to each other. Most people however can not deal well with dimensionality greater than four. Now we get many representations of 3 dimensional data in 2 dimensions. In addition 3D devices also becoming popular. It is probable with advances in 3D technology that people will arrive at ways of explaining a few more dimensions in data sets - such an education could result in more requests for representing and sorting of data in higher dimensional multi spaces. Nevertheless we are not even ready for 3D displays, we need convincing high speed 3D rendering. Ideally we would have holographic displays [2]. In a decade we will need to represent a large volume of data as a hypercube with interaction such as rotations and cutting using wands (or game controllers). We should consider the use of sound for data sets - this could be particularly useful for outlier detection. How else might we use sound ? How can we interact with terabyte scale archives? Hassan & Fluke (2011) provide a survey of visualisation on the large scale.

4 Virtual machines and light clients

Many applications are going virtual - many things running on our phones or browsers are just light front ends to functionality running on a server somewhere. On the server end, virtualization is used for load balancing which may also be needed on future archives depending on the weight of the application. Eventually everything must be in the *browser* or device and what is running should probably not require any big download. One should not be hijacked away to some other interface (such as Java applets do now) - phone applications maintain the phone interface (gestures etc).

A second big point for virtualization is to make a much richer environment available to archive users. CANFARGaudet et al. (2010) and SKA[3] are already starting on this. As we build the complex applications we will need more and faster access to the data - needing to run it in the data centre. Virtualization is only one way to do this, the requirement is to run a distributed application accessing an entire archive probably without pulling the entire dataset over the net. This will require an API and development environment for creating the distributed applications. A full container system which could *apply* an application to the entire archive for the user would be excellent.

[2] Sergei Klioner likes to show clips from star wars with holographic maps http://www.youtube.com/watch?v=LEoNA_5RACc

[3] http://www.cyberska.org/

4.1 Astro applications store and social networking

Once we have the fancy API and can make applications to run in the archive it would be interesting to publish them in an *applications store* having a ranking facility with user reviews, discussion forums - no need to develop anything as this already exists, all we have to do is to tie in. Ideally this would all be available with a single existing sign on - many sites now offer login with Yahoo or Facebook IDs - why not an archive? Imagine having a community building AIPS++ into the archive - it could be done in this manner at no cost to the archive providers. Also the interface would have to be *widget* oriented allowing the user to full customize the screen for their preferences.

5 Outreach and simplicity

We are far behind in the aspect of outreach in Europe. Ideally the archive would be out reach oriented - simple enough for school children to use at the outset with more advanced features available for the professionals. Google topped the world with a single entry area and a button. Ideally of course the user could customize the system for personal taste by choosing a range of applications or widgets which they wish to use regularly. Also simpler interfaces run better on mobile devices and mobile is the way forward. In addition there should be lessons and activities centred around the archive - SDSS has several good examples built in the same website the Sloan astronomers use.

On this note for proper connection to people the system must be properly multi-lingual from the ground up, including menus etc. Also since apparently China will overtake the USA as the i world leader in Science in 2013 - so perhaps it should be in Chinese.

6 Reprocessing

Catalogues moved online from paper recently - the first offerings just allowing searching. The virtual observatory brought some new possibilities for interoperation and cross matching but it is still rather clunky and to some extent bogged down by democracy. ESO have been offering on the fly reprocessing of image data for some years now. We need to provide more sophisticated access mechanisms. This is covered to some extent in Sect. 4 but for reprocessing original raw data must be stored and reprocessing must be a simple to execute task. However local storage for intermediate datasets will be needed. This is done on CasJobs O'Mullane et al. (2005). All of this implies having software available - sourceforge (or equivalent) is a logical place to build an open source community to support open sourced archive and processing software. This is very challenging for something like Gaia astrometry.

6.1 Living archive

What if one have several observations of an object and wishes to combine it with other archive observations. Then, you could download the archive observations and do the processing but if the result is interesting, say an improved periodicity or better orbital reconstruction, would it not be good to add the observations to the archive? Anthony Brown calls this the living archive. Obviously this is a logical advancement when the archive is no longer bound by paper.

6.2 Archive as a model

Many groups are involved in modeling and wish to compare models to archives, see Hogg & Lang (2008) describing the idea of the catalogue as a model of the data. For us this could have a profound impact on our data processing, making a challenge to think more in terms of model fitting and residuals. Gaia, to some extent, does the processing in a compatible manner for this sort of model comparison in which we do fit a model to the data - allowing someone else to fit their model however that would be a very *difficult* job.

7 Pay on demand

Sophisticated archives may require sophisticated funding - if it was all virtualized one could consider allowing people to purchase time on virtual machines, from Amazon for example, to do the processing. Another approach would be for the archive to be a facility where big user or complex projects would have to apply for time on the system. The latter is how current super computer centres work. The former has the advantage of not requiring the archive to become a supercomputing centre but has the problem of co-location of data to processors e.g., the entire system would have to be in Amazon Ireland for it to work.

8 Conclusion

A few ideas have been briefly outlined for future archives. As always these are exciting times. Who knows how archives will look like in a decade or two. At the outset of Hipparcos no one considered the entire catalogue could be shipped on CDROMs - the Gaia white book considers the possibility of distributing the catalogue on a dedicated device. To break through to the next level we may need to consider also unstructured data in publications with analytics built in; we need to harness the latest user interfaces and we need to scale to make it fast. In the end we must remain nimble enough to take advantage of new technology and should finally consider leaving the paper behind.

9 References

Gaudet, S., Hill, N., Armstrong, P., et al., 2010, In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 7740 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ADS Link

Hassan, A., Fluke, C.J., 2011, ArXiv e-prints, ADS Link

Hogg, D.W., Lang, D., 2008, In: American Institute of Physics Conference Series, vol. 1082 of American Institute of Physics Conference Series, 331–338, ADS Link

O'Mullane, W., Li, N., Nieto-Santisteban, M., et al., 2005, In: International Conference on Web Services,
Also MS technote `http://arxiv.org/pdf/cs.DC/0502072`

Perryman, M., 2009, *Astronomical Applications of Astrometry: Ten Years of Exploitation of the Hipparcos Satellite Data*, Cambridge University Press

A Acronyms

The following is a complete list of acronyms used in this document. The following table has been generated from the on-line Gaia acronym list:

Acronym	**Description**
API	Application Programming Interface
CDROM	Compact Disc Read-Only Memory (also known as CD-ROM)
ESO	European Southern Observatory
GST	Gaia Science Team
ID	Identifier (Identification)
SDSS	Sloan Digital Sky Survey
USA	United States of America