Project STORE: Astronomy Report

Sayeed Choudhury (Johns Hopkins), Robert Hanisch (Space Telescope Institute) and Rowena Stewart (University of Edinburgh)

Contents

Summary

In many ways, digital astronomy is at the forefront of issues related to data curation, given the existing experience with generating large amounts of data in raw form, and significant quantities of derived data in processed form. Additionally, astronomers have agreed upon a set of standards and web services for accessing, organizing and disseminating data. In the United States, the international Virtual Observatory effort is often cited as the archetypal example for cyberinfrastructure-related discussions. Astronomy data is “unconstrained” in the sense that it does not contain the same privacy, legal, commercial, etc. parameters of other scientific disciplines. This characteristic enables astronomers, and librarians, to build systems in an open manner.

1) Apart from being a condition of use of source repositories, the culture in astronomy is strong for citing source data in publications. Links from output to source repositories may be more useful than vice versa. The main value for accessing data in this manner would be value to the research community, to validate results, to identify specific astronomical objects of interest, or to identify collaborative opportunities.

2) Researchers are happy for their (source) data to be used as long as it is credited and, where publicly funded, there is an obligation for it to be made so anyway after a proprietary period of usually 6 to18 months (during which time data is restricted to project team members).

3) ArXiv.org and NASA-ADS are the main A&I database and output repositories used. The Virtual Observatory team and Sheridan Libraries at Johns Hopkins are working with the University of Chicago Press to consider output repository support at the time of article submission, especially as it supports preservation of derived data cited within publications.

4) There are facilities to link source to output data in operation, e.g. CDS's Simbad but they are not comprehensive and one interviewee mentioned his work on improving the linking.

5) Source repositories like being able to monitor how much they are used, especially if metrics for use might help gather additional funding or support.

6) Astronomers should define standard methods to refer to same objects when viewed through different spectra, including the provenance or annotations with certain data (or analyses of data) are deposited into output repositories. Additional metadata through automated mechanisms (e.g., telescope directly records weather conditions) would also be useful.

7) Astronomers would not seek help from librarians or informational professionals with information seeking or navigating, but rather for assistance with metadata and preservation matters related to datasets.

Survey

This report provides an overview and details from the survey and interviews with astronomers through Project STORE. The information that was collected first within the survey related to institutional affiliation, professional identity, and discipline. Given the connection to Johns Hopkins University, astronomers from both the UK and US responded to the survey. The survey included responses from sixty-four astronomers at the following thirty-one institutions:

- Advanced Camera for Surveys @ the Johns Hopkins University
- Anglo-Australian Observatory
- Astrophysics Group Physics Department Imperial College London
- Dominion Astrophysical Observatory, Canada
- Durham University
- European Space Agency
- Institute for Astronomy Edinburgh
- Institute for Astronomy Royal Observatory
- Institute of Cosmology and Gravitation (ICG), Mercantile House, Hampshire Terrace, Univ. of Portsmouth, Portsmouth, PO1 2EG, UK,
- Jodrell bank Observatory, The University of Manchester
- Johns Hopkins University
- NASA Goddard Space Flight Center
- National Radio Astronomy Observatory
- Naval Research Laboratory Washington, DC USA
- Open University
- PPARC. UK astronomy technology Centre
- Radio and Space Plasma Physics Group, Leicester University
- Space Telescope Science Institute
- Space Telescope Science Institute Baltimore, Maryland, USA
- United States Naval Observatory
- University of Leicester
- University College London
- University of Cambridge
- University of Edinburgh
- University of Edinburgh
- University of Exeter
- University of Hertfordshire
- University of Leicester
- University of Nottingham
- University of Sheffield
- University of Sussex

There were multiple respondents from the University of Edinburgh, Johns Hopkins University, Open University, Space Telescope Institute, University of Leicester, University of Sheffield, University of Nottingham, and the University of Sussex. Given that astronomers at the University of Edinburgh and Johns Hopkins University sent specific email messages to their colleagues, it is not surprising that these institutions were especially well represented in the survey. That is, the larger number of respondents most probably reflects the impact of personal communication rather than a special interest in the survey topics. The topics examined by Project STORE almost certainly have widespread for the astronomy community throughout the UK and US.

Survey respondents identified themselves with the following distribution of roles:

University Academic Staff	48%
University Research Assistant	14%
Postundergraduate Student	14%
Contracting Researcher	5%
Independent Researcher	2%
Other	17%

Within the overall discipline of astronomy, the respondents identified the following (unique) main fields of interest:

- Astronomy Stellar Evolution
- Astronomy - in particular planetary systems formation and evolution
- Astronomy & astrophysics
- Astronomy Astrophysics Galaxies
- Astronomy Astrophysics Galaxy evolution Star formation
- Astronomy Astrophysics Scientific Databases
- Astronomy Cosmology Numerical Simulation
- Astronomy, astrophysics, astrobiology
- Astronomy, large databases, astronomical instrumentation, survey astronomy, galaxies, cosmology
- Astronomy: interstellar medium; star formation
- Astrophysics
- Astrophysics (Observational and computational)
- Astrophysics and Space Science
- Clusters of galaxies radio astronomy
- Computers, Astronomy, Physics
- Cosmology
- Cosmology - large scale structure - statistical descriptions of large datasets
- Cosmology; galaxy formation and large-scale clustering
- Data Curation, Astronomical Archiving
- Data management data discovery data access multi-wavelength data integration
- Dust, ISM, Star formation
- Extragalactic astronomy
- Galactic astrophysics
- Galaxy formation, AGN
- High energy astrophysics

- Interstellar Medium Stellar Populations Cosmology Supernovae and Supernova Remnants
- Large Databases
- Observational Astronomy
- Observational astrophysics Space instrumentation
- Physics
- Physics (astronomy)
- Physics, Applied Mathematics
- Polymer physics
- Population synthesis, stellar winds, galaxy evolution
- Proto-planetary disks stellar atmospheres planet formation chemistry atomic data
- Solar Physics Plasma Physics
- Solar System astronomy. Comets, asteroids.
- Solar Terrestrial Physics
- Spectroscopy in the FUV: Hot Stars in Globular Clusters Emission from the Diffuse Interstellar Medium
- Star formation in spiral galaxies. Supernovae.
- Stellar astrophysics binary stars Telluric ozone
- Stellar spectroscopy Stellar Evolution
- Wide field astronomical surveys Databases Information technologies

Two of the respondents noted that it would have been helpful to include the RAE categories on the survey itself.

The Need for Linking Repositories

This section of the survey featured brief definitions of source and output repositories. Respondents considered the following two questions, and provided their responses as follows:

"Source repositories contain primary research data. If a standard feature of such repositories was the ability to identify and link to the publications that had been developed from these data, how advantageous would you find it?"

Significant advantage to my work	45%
Useful but not of major significance	34%
Interesting but not particularly useful	13%
Of no interest to me	2%
Not sure at this point	3%
Other	3%

Among the free form comments, one respondent stated, "I would find it a fairly dangerous development." This assertion may relate to another respondent's comment that "Data are objective, whereas interpretations are subjective and the two should only be collated with great caution since some people (especially students) tend to give as much credence to a fashionable finding as to the actual data." More than one respondent indicated that such a service exists through the ADS at Harvard and SIMBAD at Strasbourg. These individuals stated that the service is helpful in tracking down literature, but the process for generating such links could be improved and automated. One respondent whose primary field of work relates to storing and archiving source repositories (perhaps obviously) indicated that s/he has less interest in output repositories, but it might still be useful to know how the data is being used in publications.

The next question related to links from publications to the primary source data:

"How advantageous to you would it be if it were possible to go directly from within an online publication (electronic journal article or other text) to the primary source data from which that publication was developed?"

The responses were classified as follows:

Significant advantage to my work	36%
Useful but not of major significance	55%
Interesting but not particularly useful	6%
Of no interest to me	0%
Not sure at this point	0%
Other	2%

Several respondents indicated again that such a service exists through ADS at Harvard. One respondent stated that this service "would be useful only in cases (e.g. optical astronomy) where the primary data is not currently kept in public archives." Another respondent raised the important point that such utility allows astronomers to examine and (perhaps) verify results with "controversial" results.

Research Data and Source Repositories

This section of the survey addressed data formats, source repositories, metadata and reasons for modes of access to others' research datasets. The first question in this section addressed electronic source data:

"What kinds of electronic source data do you produce? (select all that apply)"

From the list of choices, the respondents identified the following subset:

File format	**Number of responses**
Raw Data	46
Drawings, Plots	41
Images	40
Databases	37
Text-based files	33
Statistical data	29
Spectra	22
Synthetic data	18
Instrument data	14
Photographs	5
Video	2
Geophysical data	1
Plans, Maps	1
Quantitative and Qualitative details of the data	1
Quantitative questionnaire data	1
Remote sensing	1
Telemetry	1

One of the respondents also mentioned “physical models” or “theoretical data” as another file format.

The next question focused on the formats for source data:

“In what formats are these source data held? (select all that apply)”

From the list of choices, the respondents identified the following subset:

Format	**Number of Responses**
Flat files (e.g., FITS)	50
Plain text (.txt)	41
Image files (e.g., .jpg, .tif, .bmp, .gif)	40
Tables/catalogues	35
Portable document format (.pdf)	32
Hypertext mark-up language (HTML)	18
Databases (e.g., Access, MySQL)	16
Spreadsheets (e.g., Excel/.xls)	16
Word processed files (e.g., Word/.doc)	14
Extensible mark-up language (XML)	12
Statistical software	1

One respondent pointed out that FITS is not a flat file. Three respondents identified formats outside the set of choices provided: FORTRAN binary files, HDF proprietary format, and IDL database format. One respondent offered a strong opinion about proprietary formats, stating "God preserve us from idiots who archive data in proprietary commercial formats (excel spreadsheets and MS-word documents)!"

The next question focused on the idea of combinations of data formats:

"Are the data you generate sometimes a combination or group of different data formats (see MoreInfo)?"

The respondents offered the following distributions of responses:

Often	42%
Sometimes	31%
Rarely	16%
Never	6%
Potentially	5%
Other (please specify)	0%

One respondent pointed out that one of the standard publishing format is Tex/LaTeX with embedded postscript for graphics.

The next question raised the topic of source repositories:

"To which source repositories do you submit your data? (select all that apply)"

The responses reflect a diversity of approaches to this question, perhaps in part to the difference practices of UK-based and US-based astronomers. The responses included:

Source Repository	**Number of Responses**
None	28
CDS (SIMBAD and Vizier)	13
NASA-IPAC Extragalactic database (NED)	6
SuperCOSMOS	5
Astronomical Journals	4
Multimission Archive at Space Telescope (MAST)	4
Virtual Observatory (VO)	4
ADS	3
ArXiv	3
LEDAS	2
NASA Astrophysics Archives (e.g., adc.gsfc.nasa.gov)	2
Very Large Array (VLA) Archive	2
Advanced Camera for Survey Science Archive	1
CADC	1

CERN	1
European Southern Observatory Archive	1
Merlin archives	1
NOAO Science Archive	1
UK NGS	1
University website	1
VLBI (EVN) Archives	1
WFCAM Science Archive (WSA)	1
www.swift.ac.uk	1

It is worth noting that some respondents may not have been aware that they are using the Virtual Observatory (VO). For example, MAST is part of the VO, yet more than one respondent mentioned it without citing the VO.

The next question built upon the previous one by asking how often respondents submitted data to these aforementioned repositories:

"How often have you submitted data to any of these source repositories? (Tick any that are applicable)"

Noting (understandably) that the respondents stated "never" for repositories that they did not choose from the previous list, the responses below describe cases that include positive responses:

CERN

Frequency	**Number of Responses**
On several occasions	1
Frequently	0
Once	0
Never	39
Never, but I am intending to do so soon	1

SuperCOSMOS

Frequency	**Number of Responses**
On several occasions	2
Frequently	2
Once	0
Never	40
Never, but I am intending to do so soon	1

UK Data Archive

Frequency	Number of Responses
On several occasions	0
Frequently	0
Once	0
Never	0
Never, but I am intending to do so soon	3

Other (which ostensibly includes the entire range of repositories aside from the three mentioned above)

Frequency	Number of Responses
On several occasions	13
Frequently	15
Once	0
Never	22
Never, but I am intending to do so soon	4

The next three questions related to metadata:

"By selecting the following options, please would you indicate what types of metadata you consider it important to assign to your data. The metadata given in the following list are generic and you can use the 'Other' option to enter more discipline-specific terms if that is appropriate."

Metadata type	Number of responses
Author/data creator name(s)	54
Title of data set	50
Date (e.g., of data creation)	49
Format (e.g., PDF or HTML)	48
Project title	48
Project description	45
Subject keywords	40
Project reference number/identifiers	31
Publisher	21
Dates of project	21
Other (please specify)	15
Funding source	8

The fifteen respondents who cited "other" metadata types mentioned the following items:

- Description of the instrument operating mode and a detailed format description such that others could process the data
- Description of data structure
- Detailed format information for binary files
- In astronomical images it is essential to have positional information, calibration information, instrument set up information
- Lots of astronomical metadata, e.g. celestial object, position, observation date, data reduction software versions, etc.
- Over three hundred and fifty other details containing data detailing the instrument used, instrument operating conditions, atmospheric conditions, light conditions, error margins, data pipeline used, data pipeline operating conditions, filter and reprocessing information. More metadata is created but stored in the database system rather than with the individual files (though there are links from the file to the database and visa versa).
- Processing method and version
- Reference of published paper connected to data
- Relevant field/sub-fields
- Specifics of the data processing steps used in creating the product
- Summary of input parameters of run which produced the data
- Telescope, instrument
- Various astronomical parameters

Two of the other respondents also made these additional comments:

- "I do not think one should include publications under 'data'. It is important to recognize AND PRESERVE the fundamental difference between them."
- "I think there should be a core mandatory list and then an optional one. The latter could be as long and comprehensive as people like to be e.g. link to ADS paper, where data has been published; data source; software + version used in the project."

The next question addressed the stage of metadata assignment:
"At what stage are metadata assigned to your data? (select all that apply)"

The respondents offered the following responses:

Stages	**Number of responses**
During file saving	25
Prior to data creation	20
When submitting data to the repository	18
As part of the indexing process for source data files	15
I am not certain of the stage at which metadata are assigned	13
After submission of my data to the repository	3
No metadata are assigned	3
Other (please specify)	0

Two of the respondents pointed out that metadata are assigned at multiple stages. For example:

- "Basic imstrumentation [*sic*] and pointing information is added at data creation. When the images are reduced and calibrated additional metadata is included. If images are catalogued additional information to identify catalogues and some information from the catalogues is put into the image metadata. When it is archived additional metadata is added."
- "Meta data is assigned at various stages, as relevant. For example the observers names are assigned at the telescope, the data version and release date are assigned later."

The final metadata question addressed the question of who assigns metadata:
"Who assigns metadata to your research data? (select all that apply)"

Role	**Number of responses**
I decide which terms to use and I assign them	27
Metadata are generated automatically	23
Metadata are assigned by the repository administrators	12
It is not known who assigns metadata	10
Research colleague(s) assign metadata on the team's behalf	5
Research support staff assign metadata on the team's behalf	5
Other (please specify)	5
Metadata are assigned by library/information services staff	2

Respondents indicated that astronomical instruments often generate metadata automatically, especially noting that the volume of data makes it impractical to enter metadata manually. They also indicated that there are evolving standards for metadata within astronomy, which are used whenever possible. For other types of metadata, project or mission teams typically make decisions regarding specific terms.

The next two questions on the survey related to access by others to research data:

"Why might you wish to access the research data generated by other research programmes? (select all that apply)"

Reason	**Number of responses**
To access data that are useful or necessary to my research	57
To test the uniqueness and validity of my research objectives	31
To understand the broader context and orientation of my research	28
To test the uniqueness and validity of their research objectives	19
To identify useful contacts	11
Other	1

The only “other” comment was “to increase the size and completeness of the overall dataset” with the following example:

- “For a recent small spectroscopic survey, we only targeted galaxies that had not already had spectra taken in other surveys, so we used the data from these surveys in our final catalogue and in our target selection.”

The final question in this section of the survey was:

“How would you normally access the research data of other researchers? (select all that apply)”

Mode of access	**Number of responses**
Through online access to source repositories	51
By access to networked file servers at other institutions	32
By access to networked file servers at my own institution	22
Through the exchange of data held on portable media	16
Other	9
I do not normally access others’ research data	4

Of the nine “other” responses, seven respondents mentioned email (though one indicated that s/he would seek it online first, and only use email afterwards). Another respondent indicated that s/he would check the data archive’s website. Two respondents referred to “analog” options with one stating “by reading of a graph” and other stating “by manually retrieving historic data (photographic plates) from archives and digitizing them myself.” One respondent mentioned the Virtual Observatory (e.g., MAST at http://archive.stsci.edu)

The Accessibility and Sharing of Primary Research Data

This section with five questions focused on respondents’ patterns or preferences for sharing their own research data. The first question focused on measures to make data available:

“What measures to you use to make your own research data available?”

Measures	**Number of responses**
By the provision of a publicised URL	40
Data are distributed via e-mail	38
Through a source repository	29
Via a publisher	16
By the allocation of passwords to network drives or data files	15
Through the exchange of portable media	15
Data are posted or passed by hand in printed format	8
Other	4
I undertake no measures to make my research data available	3

Of the four "other" measures listed, respondents mentioned an *unpublicised* URL, an FTP server, "soon via the CADC", via the grid infrastructure (AstroGrid), and through web-based visualisation programs.

The next two questions addressed factors that would encourage or discourage sharing of research data:

"What factors would ***encourage*** *you to share your research data?" (select all that apply)*

Factors	**Number of responses**
Potential benefits to the research community	56
Enabling collaboration and contributions by others	51
Improved visibility for my research	50
Demonstrable benefit to my research profile	38
Requirement of funding body/condition of funding	33
Improved level of validation for my research findings	31
Demonstrable benefit to my institution	30
Potential benefits to society	20
Other	1

The "other" response expressed "help with necessary software" and two respondents stated that they either already shared data or did not need encouragement to do so.

"What factors would ***discourage*** *you from sharing your research data? (select all that apply)"*

Factors	**Number of responses**
The time/effort required to enable sharing	41
Risk of premature broadcast of research findings	38
Risk of diversion from principal objectives through the generation of additional work	28
The threat of loss of ownership	21
Risks to an established research niche	18
Increased competition for funding	15
Subversion of intellectual property rights, including copyright	11
Consideration of data protection and other confidentiality issues	10
Ethical constraints relating to my research	3
Risk of commercialisation opportunities	3
Other	

Two respondents mentioned that s/he share data regardless of any of the factors mentioned above, and one specifically mentioned that none of these factors would discourage her/him. One respondent indicated that "risk of incorrect interpretation of data" is a discouraging factor.

The next two questions in the survey focused on the types of restrictions applied to data that might be shared. The first question noted formal restrictions:

"Normally, what kind of formal restrictions ***do you apply*** *to your research data? (select all that apply)"*

Restrictions	**Number of responses**
Individual enquiries/requests for access are judged on their merits	27
No formal restrictions	23
Restricted to immediate research team/programme members	19
Time related embargoes	19
Data is flagged confidential/commercial-in-confidence (or other caveat), for authorised access only	2
Other	0

One respondent indicated that funding bodies for astronomy support a period of "proprietary" access to the data during which time team members (only) have access to provide an opportunity for research, analysis, publication, etc. Following this period, data are expected to be publicly available.

"What measures do you ***normally use*** *to control access to your data by others? (select all that apply)"*

Measures	**Number of responses**
No access control – there is open access	28
Authentication of ID and password for online access	22
Storage of data on a private network/intranet	15
The specific operational terms and conditions of the source repository	10
Storage of data on standalone computers	6
Maintenance of an approved list/directory of data users	4
Reference of data requests to a review authority	3
Validation of data users by clicking on an e-mailed URL	3
Other	0

Output Repositories

This portion of the survey focused on output repositories defined as "those in which research publications are deposited." The first three questions focused on the uses of output repositories.

"Which kind of output repositories do you use to find and retrieve ***information for use in your research****? Examples of output repositories are given at MoreInfo. (select all that apply)"*

Type	Number of responses
Institutional	36
Publisher	30
Discipline	23
None	4
Other	3

The “other” responses included NED, ADS, astro-ph, and telescope-based repositories such as MAST, NOAO, NSA.

“Which kind of output repositories do you use to find and retrieve information ***for use in teaching****? (select all that apply)”*

Type	Number of responses
None	24
Institutional	22
Publisher	17
Discipline	12
Other	3

The “other” responses include ADS, Google, webpages, and others’ Powerpoint files. Two respondents indicated that they do not teach.

“In which output repositories do you deposit your research publications? (select all that apply)”

Type	Number of responses
Publisher	43
Discipline	33
Institutional	30
None	2
Other	1

The one “other” response was “alongside our source repository.” The next question of the survey focused on the routes for accessing output repositories:

“What are your normal or preferred routes of the contents of output repositories? (select all that apply)”

Routes	Number of responses
Via a known repository’s URL	53

Directly through a specific journal's website	31
From an internet search engine (e.g., Google)	28
Through an author's personal webpage	23
Via a library catalogue that links directly to an article in a repository	21
Through a publisher's online service (e.g., ScienceDirect)	20
From a link provided in an e-mail, CD-rom, USB drive, etc.	12
Via an Open URL resolver	8
Through a subject portal service (e.g., Entrez)	4
Via a library subject page	3
I have no normal or preferred routes	2
Other	2

The two "other" responses both cited ADS as the route. The next question in the survey focused on searching:

"What level of searching do you normally find sufficient when using an output repository?"

Level of searching	**Percentage**
Advanced, using a range of fields and identifiers	25%
Employing Boolean logic	9%
Simple - e.g. author, title, keyword, date	61%
Using a subject thesaurus or subject headings	2%
No Preference	3%

Six respondents offered comments of the following nature:

- "The NASA Astrophysics Data System and the arXiv.org preprint archive are sufficient."
- "Often, searching on a target (the name of a star) returns links to articles that are not about the target, but only mention it in passing. It would be helpful to be able to restrict the list to articles for which the target is the/a principal subject of the article."
- "It would be useful to be able to search output repositories on astronomical source name or position in the sky."
- "When using science citation index, it would be useful to be able to eliminate results which were in totally unrelated fields."
- "I'm quite happy with the provisions in astrophysics."
- "Google-style smart searching (i.e. Google scholar)"

The next two questions focused on the type of information support that astronomers might receive:

"Do you receive support and/or guidance in your use of output repositories? (This need not take the form of personal support from someone else but could be online prompts, links and advice from within the repositories themselves)"

Type of support	**Number of responses**
Documentary support	19
Repository-enabled support	19
No support	14
Unknown	6
Personal support through an intermediary	5
Other	1

The “other” response was "through contacts with colleagues” and one of the respondents who indicated “no support”, added the statement “it isn’t rocket science.”

“What assistance in your use of repositories is provided by a librarian or other knowledge management support? (select all that apply)”

Type of assistance	**Number of responses**
Unknown	31
Online or telephone help	10
Provision of documentation (guidance notes, fact sheets, etc.)	10
Assistance with conduct of searches	9
None	5
Full intermediary service (e.g. the conduct of searches and organisation of results)	3
Assistance with the structuring of specific searches	1
Other	1

The “other” response was “advisement of new repositories, and institutional access arrangements.” Two of the respondents who indicated “none” also added that they did not need assistance, or that they considered learning about the use to be their responsibility.

And finally…

The final question of the survey included two parts, both of which were focused on source repositories:

“Having considered your current use of both source and output repositories, and the potential relationships between the two, what functionality if any do you consider is missing from the ***source repositories*** *that you have used?”*

Respondents offered a range of responses including:

- Not sure at this time
- None (six responses)

- None from my own source repositories, they are more complete than most repositories I have come across
- Current functionality works for me (two responses)
- They currently already have functionality that exceeds my expertise
- Their general incompleteness - but that is not a criticism of a functionalilty [*sic*]
- Calibrated data
- The linking between different source repositories is the main issues. Using information at different wavelengths is a difficult task. Systems such as Astrogrid and the Supercosmos Science Archive are beginning to tackle this issue
- Finer-grain control of what data I wish to extract
- Easy links between the two as part of an beginning-to-end framework that allows the tracking of source data through its entire path to publication
- Stronger links to the output repositories. (2) Semantic search capabilities. (3) Metadata navigation capabilities
- Simple connection -> output
- Linking data for a particular object/source/type of object with relevant publications (and sometime vice versa)
- Finding pubs and data relevant to a specific need. Also, finding pubs that cite a previous specific paper.
- The most annoying thing about MAST is that it uses SIMBAD to resolve target names. But SIMBAD often fails to recognize the common names of astronomical objects. It SIMBAD were smarter, it would make using MAST (and the ADS) much easier.
- Easy access to advanced search capabilities of the repository
- Links to related publications
- Proper links to derived data and publications
- Better links to output repositories and better links between the different source repositories - the latter is being addressed through the development of the Virtual Observatory, but the former has not to date.
- More links are required. Ability to reduce raw data on-the-fly.
- I find in general that the quota allowances are too small for downloading large amounts of data
- Still do not cross-reference some of the larger databases in my field, but I am aware that work in this direction is advancing quickly.
- Very happy with what we have in astronomy via Vizier, NED and Simbad. Please don't mess with them for the sake of some aesthetic ""global model.""
- Common formats (but this is being addressed by the AstroGrid project)

"And what functionality of any do you consider is missing from output repositories that you have used?"

Respondents offered the following responses:

- None (three responses)
- Not much – am quite impressed by level of service already offered within my field
- Not sure – they seem pretty good
- Most functionality is extant

- Very happy with what we have in astronomy via arXiv/astro-ph and ADS. Please don't mess with them for the sake of some aesthetic "global model"
- Storage of data, tables, and animated illustrations in machine-readable format to accompany papers
- Stronger links to the source repositories. (2) Semantic search capabilities. (3) Metadata navigation capabilities.
- Full links to referenced articles
- Links to source data
- Simple connection -> source
- Full search capabilities and (in some cases) the full publication (rather than just information on where it can be found)
- I like ADS's interface, but would like to do searches on previous search results to narrow selection
- Don't know of any that are of any use. A single, online, comprehensive index of published research articles would be useful
- See my complaint above about articles that only mention a target, but are still returned by search engines like ADS or Simbad
- While these things are there on some level, it is hard to navigate. For example, it should be simple to find date of publication for a given piece of data. Instead it is trivial to find the date of publication of some data from that particular observing program, but not for that specific piece of data
- The choice of output format: if all could provide the same set of output formats, it would be much easier to use!
- Translate tables automatically in machine readable files; get raw data in tabular form from figures

Interviews and Workshop

As part of the Project STORE work plan, the University of Edinburgh and Johns Hopkins University worked collaboratively to further explore and augment the findings from the survey through more detailed conversations in the form of interviews. At the University of Edinburgh, Rowena Stewart conducted five interviews with UK-based astronomers through one-on-one conversations. At Johns Hopkins University, Sayeed Choudhury and Robert Hanisch co-led a workshop that convened US-based astronomers and a representative from the University of Chicago Library.

UK-based Astronomers

StOre interview – Royal Observatory Edinburgh (ROE), Thursday 22nd June 2006 (Ed.no.1)

Dr Bob Mann
Position: Lecturer
Group: Astronomy
Room: C14 R.O.E.
Tel: 0131 668 8338
Email: rgm@roe.ac.uk (best method of communication)

Background
The interviewee works in the Wide Field Astronomy Unit (WFAU - http://www.roe.ac.uk/ifa/wfau/) which is involved in archiving extra large homogeneous sky surveys. The ROE has been involved in this for decades and in 1998 the responsibility moved to UoEd.

Historically, The WFAU archiving work was the curation of photographic plates. The scanning of this optical data into digital format is coming to an end. WFAU has responsibility now for archiving data from two of the largest projects generating infra-red data. Data being archived now begins in digital format and the file sizes are much larger than those generated from the scanning of the optical, photographic, plates; all the photography equalled 20TB on disc, the first of the new infra-red projects is generating 20TB per year and is due to run for 12 years. The second infra-red project is predicted to generate double the amount per year and run for 10 years.

An image analyser scans images looking for discrete objects (skies/galaxies) and generates brightness data, etc. “Standard attributes” are loaded into a database.

All astronomers tend to be only interested in accessing the database of standard attributes, only very occasionally needing access to the source data, e.g. if looking at something which is not usually picked up by image analysers.

Researchers in the University of Cambridge (Cambridge) do the data analysis of raw data and extract the standard attributes. Cambridge sends UoEd Flexible Image Transport system (FITs) files (which were started for Astronomy images and extended for other data). Also, the metadata describing an image; when made, how analysed, provenance, structural metadata on how to use the image.

UoEd loads the data and metadata into the databases, links multiple observations of particular bits of sky and is responsible for access and curation. All operations run on the data are driven using metadata records. The scale of the data means everything is automated. The aim is for the data from Cambridge to be sufficiently well described to run.

All the metadata is exposed in the standard attributes database as thought was put into how the database would be used and how to make the metadata meaningful/provide a helpful scheme.

If WFAU databases used, astronomers are asked to reference standard papers published in the literature. There is also a standard acknowledgement which lists all who worked on data and research councils etc. The standard articles have database description and online access details. Astronomers have only recently had to learn how to query databases so are only at the first stages of how to use SQL on relational databases and therefore, the standard papers describe how to interrogate database. Extra online help is provided and help email address provided. Shortening the time spent on the databases by researches is helpful for the database hosts as well as the researchers.

Data quality – understood what is wanted and researchers should know data is good enough to use.

Research Lifecycle

For an individual researcher - Use of source or standard attributes data: Image surveys are built over many years' observations. Might want to look for data in a particular area of sky, therefore might want to check if done already.

Interview Checklist

5 & 6:

Links from output to source repositories would be more useful than vice versa.

Data in the WFAU databases is low level and it is the analysis, etc which is published, therefore want to look up if anyone has discovered similar objects in our archive previously as there is not an easy way of finding everything which has used your data or of finding a particular "strange object", i.e. would like to be able to run a query on output repositories for source database used and query for unusual objects.

> *Astronomical objects' names:*
>
> Objects have colloquial names. There is also a naming system in operation which can be compared to latitude/longitude, e.g.: XMMXCS J2215-1738.
> XMMXCS - is the code for a particular project, database or archive.
> J2215 – is the RA
> -1738 - is the Dec
> The RA and Dec place an object in a particular place in the sky.
>
> Strasbourg group are leaders in the interoperability of astrophysical archives (Centrede Donnees astronomiques de Strasbourg - http://cdsweb.u-strasbg.fr/CDS.html). They archive others' data (usually) including tables extracted from archives and journals. They are working with some of the major journals to get them to link to their database. The Strasbourg database archives names of objects and maps to RA and Dec (where in sky). This is helpful as often articles use the common name for an object. The database helps to map between names for particular objects.
>
> Historically, astronomy is split by spectra and different names are assigned to the same object by astronomers workings in the different spectra (optical astronomers and radioastronomers). Researchers needing to work on both types of data for an object therefore need to find information and data on both.

7:

Main job is to archive what's produced.

8:

Source data is held in FITs mainly. Not hierarchical.

9:
WFAU recognises researchers may want data in different formats so as they can use different tools on it and so the project makes data available in these different formats.

10:
One of the major trends of the past 20 years is that the understanding of objects requires data over the whole spectrum – multi-wave length astronomy. Therefore users want bits of data from different archives, e.g. UoEd specialises in optical and infra-red, Leicester in X-Ray, Manchester in radio data.

The split will be maintained as expertise in the techniques for each spectra is still required, e.g. UoEd has to recalibrate Cambridge data because instruments change with time and have to adjust the archive data accordingly and this needs expertise.

11:
Each database has its own individual interface but major initiatives to build the Virtual Observatory aim to make all disparate archives interoperable and introduce automation, e.g. more standardised description of data and more standardised access to data such that an astronomer would have a piece of code which would query all archives and retrieve relevant information.

12:
SuperCOSMOS as a WFAU member
Individual researcher – Virtual Observatory will make it easier for individual astronomers to populate source repositories. The move is to have all data in archives so can be used by others and from individual astronomers making own observations to telescope time being used for collection of homogeneous data which can be used by others. Data collected by an individual is his/her property and released when they see fit.

13:
Metadata assigned to the WFAU database is fully exposed and users can use everything WFAU knows about the data. Metadata includes: Area of sky, Spectrum region, Exposure length of image, discovery of data…100s of columns.

This helps with e.g., equipment sometimes has odd behaviour depending on known factors, e.g. particular angle against the moon with amps switched, etc creates particular artefacts so seeing all the metadata helps identify this.

14:
Metadata is assigned at all stages

i) at the point of observation which includes data from the camera: when, where on sky, state of the instrument.

ii) Extra data at Cambridge from the image analysis

iii) UoEd add metadata if match multiple observations of sky, i.e. for the surveys discussed, observations are spread over years and one of the things of interest is objects which have moved. Also, Cambridge data is run nightly. So matching moving objects is done at UoEd and generates metadata.

> The original coded name assigned to an object should stay in use even when object moved or different/accurate instrument pinpoints RA/Dec better.

iv) As data is used, oddities are spotted and metadata needs changed (re-calibration)

15:
Metadata assigned by colleagues

16:
Access to data is online, especially as sizes are so large, preference is to have data FTP'ed onto a work station. This is starting to change as the scale of archive increases and users want to analyse large quantities of data, database owners don't want that much extracted at once. WFAU is beginning to offer online analysis.

17:
WFAU project - Sharing research data is the point of the project.
For an individual – it is in your interest to have your data re-used as your papers published describing that data will be referenced.

18:
Want to use your data to the full, but often re-use is in combination with other data you don't have or which will be taken later. Wouldn't want to be scooped on the obvious but you can't do it all yourself.

19 & 20:
Generally, after observation is made data is restricted, for a year, to individuals, groups or nationality. WFAU databases largely hold data taken on behalf of European astronomers, ESO of which most European nations are members. WFAU restricts access to ESO for a year to eighteen months which means there is a list of registered users able to use data, i.e. a list of ESO users. Users log in. After 12-18months, data released to open access.

Databases tend to impose a limit on the amount of data which can be returned by a query. Often queries returning a large number of attributes (more than 100 million) are bad/poor queries anyway.

21:
Astronomy is a small field and NASA-ADS and ArXiv.org fulfil output repository needs.

22:
It is unusual for UGs to use research data. Project work may require it and they would use ADS too.

23:
Deposits in output repository, ArXiv.org and then the journal will automatically deposit in ADS. UoEd Institute for Astronomy used to maintain its own output repository but no longer does now that ArXiv.org so widely used.

24:
ADS,etc are bookmarked

25:
Searches on abstract, titles, author, full-text, keywords

Keywords:
For ADS and ArXiv.org are author assigned from a standard list agreed by journals. As an author the interviewee knows he does this quickly and that they are not particularly descriptive.

26:
Online help

27:
No assistance sought from librarians and wouldn't expect, e.g., ADS, staff to help.

28: Missing from source data?
Strasbourg (Simbad) links into databases and ADS links too via, e.g. Strasbourg for individual objects. Links are worked out manually at Strasbourg or Pasadena (ADS) and takes a lot of effort. Therefore would like in longer run to have an automated spatial query for any object in a particular area of sky (see 11 above)

29: Identification with perceived need of links from source to output?
Want to know who's using data as helps WFAU to get continued support.
Individual researchers – other person referencing your papers. Although if the Virtual Observatory reaches stage where individuals can publish databases, not just papers, there may have to be another mechanism of referencing.

30: Linking to source data from publications – metadata problems?
Spatial indexing which is a natural means of locating data. Technical issues of implementing spatial indexing in database; queries on a range of objects in an area of sky – difficult to do over text as there are a number of different names for the same bit of sky.

31: Missing from output reps?
ADS is really good – standard user interface restricts on kind of queries so it would be nice if there were some means of expanding the range of queries on data there already. Plus links to source data.

32: Benefit from ability to associate newly deposited publications with data from which derived?
Looking to the future - for an individual publishing a database in the Virtual Observatory there would be benefit in having some standard means of having standard reference system. However, astronomers are unlikely to use anything not produced by astronomers and standard metadata, e.g. DublinCore, is not enough as need more details, e.g.) for time – time of observation, when analysed, when recalibrated, ii) no always obvious with astronomy data who are creators or owners.

33: New operations supported within an output repository – how meet your needs? Others advantageous?
Better ways of presenting what ADS has already.

The interviewee is working with UoEd School of Informatics on text mining online literature for e.g., looking for articles describing particular classes of objects. It is not easy to match strings of text due to the different names associated to objects across spectra, etc. Looked at machine learning techniques [with Ewan Klein & Clarie Grover]. Machine learning to assign automatic keywords/better keywords.

34: dataset knowledgebase – value and specific issues
How would you use user annotations in a database, e.g. for an archive of 100 million objects with 100 of one particular species, if some user found them all and publishes something which has used objects 101 and 103, might not want to add their information on these two objects to WFAU database but other users might find the information useful, therefore want to link annotated database to source database and the interviewee has done work on that with Rajenda Bose (Informatics).

WFAU databases hold low level data but might want an extra level where users can publish their annotations.

Thinking about linking all archives into the Virtual Observatory, want a means of researching what entry in optical database conveys to which in X-Ray database and this needs a lot of effort. Would like a means of matching automatically (using algorithms). Cf Strasbourg which only going from published data and doing manually and also want what is stored in databases not just in papers. The interviewee would like/envisions, within the Virtual Observatory using a standard query language to search all databases held in the Virtual Observatory which would present matches retrieved from the different databases and link to them.

35: Control necessary & access validation?
Control necessary for access to data during its proprietory period (see 19 & 20 above) and would expect that control to be by username/password.

Ultimately, it's astronomy data, therefore control does not have to be high level as there are not personal/medical/national security etc issues.

References

R. Bose, R. Mann and D. Prina-Ricotti (2006) *AstroDAS: Sharing Assertions across Astronomy Catalogues through Distributed Annotation*. In: Proceedings of the International Provenance and Annotation Workshop (IPAW'06) [http://www.ipaw.info/ipaw06/], Chicago, IL.
http://homepages.inf.ed.ac.uk/rbose/pubs/200605_IPAW/bose_ipaw06.pdf

Claire Grover, Harry Halpin, Ewan Klein, Jochen L. Leidner, Stephen Potter, Sebastian Riedel, Sally Scrutchin, and Richard Tobin (2004) *A framework for text mining services*. In: Proceedings of the Third UK e-Science Programme All Hands Meeting (AHM 2004).
http://www.iccs.informatics.ed.ac.uk/~ewan/Papers/Grover%3A2004%3AFTM.pdf

StOre interview – JCM Library, King's Buildings, University of Edinburgh, Tuesday 27th June 2006 (Ed.no.2)

Dr Ken Rice
Position: Lecturer
Group: Astronomy
Room: C15
Tel: 0131-668-8384
Email: wkmr@roe.ac.uk

Background
The interviewee's field is theoretical and computational astrophysics and he models planet and planetary system formation. He is "not an observer as such" so the data generated by his models are not ones people would particularly want to use.

Research Lifecycle
Scans ArXiv.org's Astro-ph regularly to keep up to date and when writing a grant proposal will remember a relevant paper and chase up in Astro-ph or ADS. Occasionally uses WoS as his area was not so well represented by Astro-ph or ADS but ADS is updating itself. ADS felt to be relatively complete.

For data would draw from Centrede Données astronomiques de Strasbourg's SIMBAD Astronomical Database (http://simbad.u-strasbg.fr/Simbad) or California & Carnegie Planet Search (http://exoplanets.org) . The latter lists all known information about extrasolar planets.

Would go back to publication and source data repositories from grant application stage through to post-publication presentation stages to update with more recent observations as necessary.

Interview Checklist
5 & 6:
Links to publications from source data would help to, not being an observer, get a better feel and extract data. More useful to link to source data from publications as ADS via Simbad.

7:
At the moment, make up own files, ASCII, binary. Text based files contain all the numbers relevant to a simulation. If very big, raw data stored in binary format.

8:
Held in text format or binary files of GByte size so stored on own hard drive.

9:
Does not commonly generate data in a combination or group of different data formats.

Has a basic file of all positions (masses etc of different data). Data is extracted and a position plotted. Doing something different, will extract again. Duplicates are discarded when analysis done.

10:
Theoretically it is difficult to imagine simulation data being wanted by others because, unlike in cosmology in which datasets are so big you extract the bits you want to work on, in the

interviewee's field, sets are small and used really just for running a simulation. However you may want to use others' data for the following reasons:

Possibly, if don't trust what another researcher has done you wouldn't trust the data and so want to check it

From a numerical simulation, e.g. density map, x-ray astronomers could use your data etc, therefore there is a possibility someone might want your data.

The interviewee has looked before at easily accessible spatial data, although looks at extra-solar planets standard attributes, i.e. data produced from analysis of raw data as the raw data is of no use to him.
There is a debate about the use of modelling data. The process is changing with multiprocessor machines running the simulations which therefore takes longer so can't just run your own simulation on the same data to get figures. Traditionally, data is collected to answer specific questions and when a simulation run, the need for the data and the data produced from the simulation is generally over. Now, bigger sets may mean there is more to get out of the data after used once.

11:
Has used ADS links to Simbad. Gone straight to exoplanets which relies on groups or individuals to maintain. Does contact individuals but more for output (graphs etc) rather than raw data.

12:
Does not deposit source data. Can see it happening in the next decade but there could be issues with server space and who does the uploading and file formats

13: What metadata is assigned?
Not applicable, but would like to see.

If a simple file, the first line is how many lines in the output file, how to read it, assumptions on the model, type of code, type of simulation, equation of state, type of object (but normally scaling factors).

14:
Now, for even small datasets, generate new file every time a run produces output. The alternative would be a single file from the beginning.

15:
Not applicable but would expect whoever manages (a future) repository would state what metadata should be added and researchers would add what is required/agreed to their data.

16: How do you make your research data available?
ADS links appear when a simulation relates to real data. Systems referred to are the ones linked via Simbad.

Would e-mail it out if people asked (if about to write up, may wait a while) but this has never happened for source data, usually for a figure for a presentation.

17:
Happy with the idea of sharing if it builds up collaboration. Would not share immediately after a simulation finished.

For the models themselves, the interviewee gets them from others and modifies them. Models percolate through the community via collaboration (someone's worked on a model and s/he works with someone else who then uses the model, etc) rather than folk approaching other researchers for a particular model. Percolation happens as modifications to a model take little time. If someone spent a year working on a model, they could feel differently but then others' use of it would mean their work would be referenced and their name get known.

18:
Discouraged from sharing if just about to publish. Once data used, happy to give it out.

19:
Could imagine a pre-publication embargo. The collaborative groups in which the interviewee works are small (5-6) so a restriction to just group members (cf Bob Mann interview) would be unlikely to be necessary.

20:
Actual control measure is that people have to ask. But this is only because data is not on open access for reasons given above (not thought worthwhile, small datasets)

21:
For research output reps – Astro-ph on ArXiv.org, ADS, WoS

22:
For teaching, resources include e.g. Hubble heritage site for images. From text books, etc, examples to show of ways of solving equations. [i.e. research publications not used nor their output reps searched for teaching resources]

23:
The interviewee loads pre-prints on Astro-ph on ArXiv.org. Then work deposited wherever journals put them.

Was vaguely aware of UoEd's institutional repository and would consider depositing in it because it would be the "right thing" to do. However, at previous employer, University of California, he didn't trust the download figures from their institutional repository (seemed too high compared to those from ADS). Did mention CiteBase figures as something he would trust.

24:
URLs bookmarked or on browser history. Sometimes Googles for the URL.

25:
Scans all of Astro-ph on a daily basis.

When putting together grant proposals or writing publications, he will remember an author from the scanning and search Astro-ph for the name.

If doing a trawl of the literature the interviewee will use keywords, etc but lots of hits usually result so doesn't use this method often.

26:
Online help now and again for specific problems but generally just try things.

27:
Use of librarian or some other information professional if can't get into something, e.g. Icarus online, rather than help with navigating repositories.

28: Missing from source data?
Pretty much satisfied

29: Identification with perceived need of links from source to output?
Standard attributes – a way of finding databases of standard attributes which the interviewee doesn't look for at the moment as it's complicated. If getting the basic data was more straightforward he would consider using such databases.

30: Linking to source data from publications – metadata problems?
Can't think of any.

31: Missing from output reps?
Fairly happy. They are getting more advanced by the day.

32: Benefit from ability to associate newly deposited publications with data from which derived?
Useful to find the data from publications.

33: New operations supported within an output repository – how meet your needs? Others advantageous?
Sounds ideal but seems very complicated. Might encourage you to look for "all the known information about an exoplanet" or "what if plot x vs y" for a bigger sample of data than the specific data from the simulation. Or if interested in all systems, not just particular ones.

34: dataset knowledgebase – value and specific issues
The idea of why the data is being used is simple as want simple data on which to run models. Could see how such a dataset knowledgebase as described, could throw up new ideas though.

35: Control necessary & access validation?
Data for collaborators for 6 months via u/p then OA.

Or, don't make available until finished with it. Wouldn't see need for any access control after finished using data.

StOre interview – Royal Observatory Edinburgh (ROE), Wednesday 28th June 2006 (Ed.no.3)

Dr John Davies
Position: UK Astronomy Technology Centre Staff (works for research council)
Group: Astronomy, PPARC
Room: R. 23/F
Tel: 0131 668 8348
Email: jkd@roe.ac.uk

Interview Checklist
5:
Nice to have links from publications to data but whether or not they would be used is a different matter. The interviewee collaborates with three or four others, each having a different role. His collaborators may want to look at raw data to fine tune a paper, so could see the usefulness of links.

Possibly useful too, to get from elsewhere, others data for the same object and use it in amalgamation with your own.

Could be used to check if peculiar outcomes elsewhere are real or incorrect analysis. This could take a huge amount of work for which the interviewee has no time at the moment but others could do it.

6:
Links from source data to publications would be less useful as a Google or ADS search on an object usually suffices, eg there is a particular bright object for which the errors are too large and would like to look for all the data out there for this object but never had time to.

There are other ways of getting source data; most telescopes have archives. The Virtual Observatory coming too.

7:
spectra and photometry, measures of brightness and shape, ie image files.

From the pictures, brightness data is calculated.

Therefore, raw data is useful. Before got from a single detector from which you get data on the day. Now you don't take the observations yourself. You put in a request to a telescope with the relevant parameters and on a suitable day the observations are made. In addition, the equipment is more complicated, so there are lots of ways of getting numbers, and you don't know what was actually done to get the data. Therefore, you may want to get source data and run your own fitting program.

(Images are better for source data; more useful than spectra which are more accurate).

8:
FITs

There's a new thing - Virtual Observatory table, an international standard – which will be the "new FITs".

FITs assumes spaces between data are equal but, esp in spectral data, widths are different and VO table will address this.

9:
Sometimes produce SDF files which are hierarchical (hds) files. SDF is another standard bit like FITs. Have to convert to FITs when sending out and also funding for the project (Starlink - http://www.starlink.rl.ac.uk/) has ceased and it is not being maintained so will disappear.

10:
May want to check that believe results and just possibly to extract new data, eg heard an asteroid has turned into a comet so want to check if the data confirms that.

Occasionally, for added value to look for something the original project was not looking for, esp. as asteroid changes.

11:
email to get data as rather go to a person who will know about possible wrinkles with the data, eg Tuesday's data was poor because it was cloudy or something not working just right. You don't get that kind of information from a web archive.

For a recent paper, the interviewee got asteroid spectra from links on another researcher's personal web page to which he was directed by links included in one of the researcher's papers. The interviewee collected from the page data in ASCII format, converted it to SDF and used it for his paper. He also got data on the same object from another researcher by asking directly and having the researcher email a FITs file.

12:
Doesn't deposit

Planetary data system (http://pds.jpl.nasa.gov/) – occasionally gets asked to put data up but doesn't as nothing in it for him but it also takes time as the data formats have to be altered to suit. If he was paid to do the alterations and depositing or if the archive did it themselves he would.

13:
Metadata which the telescope adds automatically; date, time, pointing position.

What's missing is:
i) weather conditions,
ii) what the observer is seeing ("how much do stars twinkle" – if the stars are extra twinkly not all their light may be captured by the instrument and this would provide different readings from what might be expected, or it might suggest there are other bodies near the star when it is just light from the star itself – N.B. this is the interviewer's attempt to translate a verbal explanation accompanied by hand demonstrations)
iii) colloquial information which would have been noted in an astronomer's journal if they had been taking the observations themselves. It's a bit difficult to put this in the telescopes archive record, eg, UK infra-red telescope observations done by other people and during the night supposed to have a thing where operators put in notes like "wobbly telescope", so that the archive has a log which is observation, note, observation, note etc but don't have the time to make these particularly meaningful.

14:
Metadata assigned when observation taken. Hds files may have entries that on such and such a date, something done. The history file is where this would be added automatically by data reduction package.

15:
Metadata is all assigned automatically by software. Could physically edit the history file/array of hds but not done in practice.

16:
Data made available by publishing in refereed journals and providing access on request. Raw data is archived by telescope operator but it is possible for the researcher for whom the observations were taken to add information if they wanted.

The researcher who has requested the observations be made receives the data on CD or can FTP from the site.

Most raw telescope data is made publicly available after a year, although the interviewee has only ever downloaded his data, ie the data he asked be collected.

17:
Encourage to share research data for the “fame”, for citations. But would share anonymously (however if interviewee were to download others’ data, he would probably want to know who had produced it).

18:
Discouraged from sharing if there was something interesting left in the data which interviewee had not yet published or if data so fundamentally unreliable it should not be used by anyone. However, in the case of the latter, it would be available from the telescope’s archive anyway.

19:
No formal restrictions applied. If the interviewee is the principle investigator, can give data to anyone.

20:
Other than not putting on an accessible site, there are no actual measures restricting access to data. Any moderately sensible personal request for data for vaguely scientific, educational etc means will be accepted.

21:
Information for research from ADS, reads appropriate journals, Googles a bit, personal contact, conference publications and telescope archive.

22:
Interviewee does not teach

23:
Deposit output repositories left to journals. PPARC has no online “institutional repository” but staff are required to keep personal web pages updated with their publications.

24:
Output repositories bookmarked.

25:
Searches output repositories by author. Know someone has written something and looking for the particular paper by author is the quickest way to get hardcopy.

26:
Very seldom goes to help file. Just tries things instead.

27:
Would ask the librarian for assistance with the catalogue. More likely to ask a colleague for help with an online resource.

28: Functionality missing from source data?
Nothing missing except perhaps weather information if available (interviewer – see also 13)

29: Identification with perceived need of links from source to output?
No real need as eg, if data interesting but never reduced and then found someone else had used it and was rewarded you would be annoyed, but at yourself as you have already had the chance to do it.

30: Linking to source data from publications – metadata problems?
None.

Information needed is so basic and encoded in raw data files. As chips get bigger and bigger, more and more data stored. Increasingly talked about is storing only the processed data rather than the raw, because of the sizes involved. Analysis data takes up 10% of the space the raw data does.

31: Functionality missing from output reps?
None

32: Benefit from ability to associate newly deposited publications with data from which derived?
Might benefit from citations.

[Icarus is a US publication which is rated less highly by RAE than Monthly Notices of the Royal Astronomical Society. UK researchers prefer to publish in MNRAS but US researchers tend to only read Icarus. Interviewee thinks links could increase the number of people reading/citing papers not published in Icarus]

33: New operations supported within an output repository – how meet your needs? Others advantageous?
Not much help to interviewee

34: dataset knowledgebase – value and specific issues
Sounds like a good idea, eg would help with the display of data “wrinkles” (see 11 above)

35: Control necessary & access validation?
Whilst request variables are in the queue for a telescope (ie observations have been requested but not taken yet) would want no-one to access. Whilst in proprietary period, data is restricted to collaborators by username/password. After a year, data on open access.

StOre interview – JCM Library, University of Edinburgh, Tuesday 4th Jul7 2006 (Ed.no.4)

Dr Johann Bryant
Position: Research Assistant
Group: Astronomy
Room:
Tel:
Email: jb@roe.ac.uk

Background
The interviewee is the Database Archive Curator for two WFAU repository systems (SuperCOSMOS Science Archive (SSA), http://surveys.roe.ac.uk/ssa/ , and WSA-WFCAM Science Archive, http://surveys.roe.ac.uk/wsa/). Interviewee also acts some of the time in a Computing Officer capacity.

Interview Checklist
5 & 6:
The usefulness, from a researcher's point of view, of links from a source repository to related publications, would depend on what the researcher is trying to do. It might be interesting or give you a source of work (although most researchers have their data generated). Would allow checking of data/results although "versioning" could cause problems

> Versioning: data from an observatory, "raw data" is taken from telescope but it is processed there to some extent, but only once. This data is sent to University of Cambridge where it is processed and sent to Edinburgh (WFAU) who process some more and quality check. Data may need to be recalibrated at a later date too. WFAU are attempting to keep all versions of processed data.
>
> Telescope version of data is not kept by WFAU, Cambridge archives on tape (tape comes from telescope) and takes off tapes each time to analyse.
>
> Astronomers do want to look at images. Images stored by Edinburgh.

For those new to research (PhD and MSc students) it would point to useful data out there that they can use.

From a source repositories point of view, links would also be a good thing but the source repository has to want to do it and it's not easy.

i) every version publicly available has to stay online permanently
ii) older versions have to be available but not instantly. Images for older processor versions, WFAU are deleting because majority of interesting science done on that which is stored now. Up to now and probably in the future, old databases have been kept though; that is the metadata is kept and kept online

Direct link back to source database means greater visibility, which means may get to host/store their data. May also mean source database gets used as store of publications details by people using the database.

Most use to the source repositories in finding source data in use which can be merged with other sources.

7:
Image data

8:
FITs.

9:
To users, provide a combination of data formats but mostly astronomers use only FITs. JPGs are provided by the source repository to give a broad display of what's in the database and for posters, not used for the science. There is a JPG for each FITs catalogue stack to give a brief overview of it. Not much scope for storing metadata in JPGs.

The repositories are considering providing "cut-outs". Astronomer can specify the co-ordinates of an area and WFAU will generate FITs with metadata for that area. Will also have references to the original images from which FITs taken.

10:
The repositories want other's infra-red data (although usually in a slightly different format) to verify their own results and also data from all other wavelengths.

UKIDSS (http://www.ukidss.org/) incorporates 5 surveys which are cross-referenced against the interviewee's own databases

11:
Interviewee, for adding to repository, will go to original site or best equivalent for others' source data.

A researcher will probably go to the original source because mentioned in other research papers, etc. Astronomers tend to be good at pointing back to sources they use and sources are pretty good at linking to repositories. UKIDSS and SSA both refer to all papers when permitted.

12:
The interviewee does not undertake scientific research. However, if he generated data in the course of his work, he would make sure he had a couple of copies but wouldn't submit it anywhere (although see 23 below). If somehow it was astronomical data, he would get it into his own repositories.

In fact, the interviewee's repositories are already providing "miniature databases" for researchers to deposit any data they get from private telescope time.

13 (14&15):
Types of metadata added are "everything we possibly can"; date and time (to a microsecond as it makes a difference), every version of software used, temperature of telescope at time of being used, directions the telescope, camera and software think the telescope is pointing…

100 keywords of metadata are recorded by the telescope comprising information about how and when an image taken and atmospheric conditions, etc. Cambridge adds approximately 200 odd keywords about the image generally and produces catalogues

> Catalogue: A catalogue is linked to stacked images (lots of images on top of each other to get a better image)

Metadata also includes comments from observer which can be added whilst observation taking place. About a fifth of data is re-taken as problems at the time and the observer will write in the reason or if there was a funny noise at the time, etc. (Images rejected/which have to be retaken are kept so can be brought back, e.g. if someone discovers something with those characteristics).The observers don't see the data. [Automatic] processors look at images to some extent, then UoEd Institute for Astronomy staff look at the images for quality control and may be able to track problems back using the observer notes. Metadata added at all these stages.

"99.9%" of metadata is added automatically.

Metadata includes error margins dependent on factors of machines etc. Many of old astronomers' notes/journals were measures of error which are now calculated automatically. With the amount of data generated it cannot be done manually. Some data requires manual assessment (e.g. to differentiate between moon glint and a star) but 80% is done without human input.

Metadata is only ever added to FITs file, never removed.

14 (and see 13 above):
Continuously – main points are at the Observatory, during processing of the image files and then at the University of Edinburgh.

15 (and see 13 above):
Automatically and by humans working on the data at all stages.

16:
Data in the interviewee's repositories is publicly accessible after proprietary period finished (18months) but depends on the Principal Investigator. For the miniature databases (see 12 above) it depends on the researcher who bought the telescope time but possibly has to be released after 10 years.

17:
Money would encourage the interviewee to share the repositories' research data. Would make data more widely available earlier if the repositories had more money to do it. More access to the repositories would require more hardware and investment.

Interviewee's repositories are looking at including a channel in MyEd [University of Edinburgh staff and student portal] but if students start using it heavily, WFAU would have problems with storage.

18:
Discouragement to sharing repository data more widely are that they are too big to distribute without more investment in hardware (image and database servers), network, power and bandwidth, physical space for a server room.

19
Those people who work on the projects, Principal Investigator and collaborators, the IfA and Cambridge group staff have access to everything.

On release version of data for 18months (or as decided by researcher for miniature databases – see 12 above) access is limited to European Southern Observatory members.

Completely public after proprietary period.

20:
Username/password over and above local network restrictions by IP address, i.e. some known spam IPs are blocked by institutions.

21:
For current awareness for technical news the interviewee uses newswebsites which include links to publications. If looking for something specific and astronomy related would look at ArXiv.org Astro-ph or Google (Yahoo getting better).

22:
Interviewee does not teach

23:
The interviewee does not generate scientific output and is not sure where he would deposit it if he did. For technical-related output, the interviewee would submit work to appropriate news websites.

24:
Interviewee's route to output repositories is through news sites and a set of bookmarks.

Some source repositories have bookmarks to output repositories.

25:
Level of search would be by Author, then keyword in title. Possibly would go further but has been using search engines for more than 10 years and so knows keywords to use.

26:
Interviewee will use manuals if in "dire need". Unless exceptionally specialised, can tell from the interface what is meant to be used. If in a field the interviewee doesn't know, he will use an obvious manual but if it doesn't turn up an FAQ or search mechanism quickly will stop. Beyond that, would use an expert or talk to a colleague.

27:
Interviewee would not use Librarian or Information Professional for help with use of repositories.

28: <u>Missing from source repositories?</u>
Interviewee would like to make "his" repositories open – would like to provide for more people but would need fairly significant resources, e.g. scale up by 100% for whole University to use [not that they would want to, as the data is specialised].

Would be a separate access level, restricting free-form queries as take too much power (the simplest queries generate large data which takes a long time to process).

The interviewee's repositories are only just old enough to contain a small amount of data past the proprietary period. However, other people's experiences of public release are that the overheads are the same as for the proprietary period. If newer data was released the overheads would increase because it would be more popular.

The interviewee's repositories are not publicised at all past infra-red astronomers. Consequently, anyone who is not an infra-red astronomer is considered to be "general public". There is an issue here for database systems because astronomers are becoming general astronomers, looking at all sky surveys, rather than just working with one wavelength hence, UoEd's Institute for Astronomy is aiming to become a large source repository

29: Identification with perceived need of links from source to output?
The interviewee's source repositories are the first point of contact for publications on their data. ESO requires researchers to acknowledge data source in ESO released papers, therefore until the general public (see 28 above, para 4) start using the data, no papers will not reference the data source. Nor will any papers not be linked to from data source.

30: Metadata difficulties?
Amount of metadata is stupendous although that's not so much of a problem.

Having links directly embedded for RA/Decs (object identifiers) and URLs not an issue as incorporating them into an already big database. However, it would be easier for users if the linking is outside the database.

31: Missing from output reps?
Nothing in particular. Although:

i) Some output repositories are missing the ability to search full-text which can be helpful.
ii) Information about the source used in a paper, unless in the paper itself (structured metadata about source data).
iii) Not good for searching about, e.g. neutron star as that term used in a lot of papers about all different kinds of things. Categorisation [controlled vocabulary] needs to be done by experts in the field.

32: Benefit from ability to associate newly deposited publications with data from which derived?
Good to track use of the source repository.

The interviewee's repositories are planning to automate a submission process for deposit and association.

33: New operations supported within an output repository – how meet your needs? Others advantageous?
Not sure for astronomy how useful showing source and output together is, unless for non-experts, due to the amount of processing needed on, and the complexity of, the data. But, in computing it would be very helpful.

34: dataset knowledgebase – value and specific issues
The concept is very good. Don't think of any use to astronomy as already done, in the repositories for which the interviewee works and other source repositories. It's done as part of them trying to make themselves viable by adding these sorts of features.

In general academe might be in competition with output repositories as technical journals, e.g. The Register (http://www.theregister.com) covers the whole of the computing industry and beyond. Within each journal there are relevant previous articles and links to other useful information from relevant bodies.

35: Control necessary & access validation?
Temporary access rights.

From computing officer side, there is reticence about allowing access to non-public material. For post-18 months access control would be by IP if lots of enquiries to free stuff (this is easier than username/password from technical side).

StOre interview – Telephone, Wednesday 5th Jul 2006 (Ed.no.5)

Dr Richard de Grijs
Position: Academic staff
Group: Astrophysics, University of Sheffield.
Tel: 0114 222 45 24
Email: R.deGrijs@sheffield.ac.uk

Research Lifecycle

Need access to data in literature and are well set up for this with data from 1800s to the present via NASA-ADS. There is also Astro-ph on ArXiv.org.

From having an idea, need raw data, so apply for time on telescope or look in telescope database and search for it. Interviewee uses Hubble Space Database Archive (http://archive.stsci.edu/hst/) especially which is a mainstay for observational astronomers.

Written papers are posted onto Astro-ph

(If there is a presentation on a paper/work done a while ago, may go back to update pictures, etc.)

Interview Checklist

5:
Very useful to link directly to publications from source data. Some do already, e.g. Hubble Space Database Archive.

CDS' Simbad (http://simbad.u-strasbg.fr/Simbad) and NED (NASA/IPAC Extragalactic Database - http://nedwww.ipac.caltech.edu/index.html) link data on objects to publications.

6:
If need output information, interviewee goes to the source data and follows links.

7:
Reduced data – tables of numbers.

Some are not suitable to publish but may still be useful.

8:
Text files

9:
Figures as well sometimes, *e.g.*, recently one of the interviewee's PhD students generated a lot of figures which the journal didn't want to host (they were held by the student/research group).

10:
People apply for observing time for specific purposes but the data could be used by someone else for completely different purposes. Interviewee thinks, based on his biologist wife's work, this differs from biology data which once used by a researcher isn't used by another.

11:
Normally access others' source data via telescope archives. If they aren't available from there, the interviewee will ask the person who produced the data. A possible advantage to getting the data from person, rather than an archive, is the person may have analysed the data that bit further.

12:
For large data tables on their own – if large data tables CDS (http://cdsweb.u-strasbg.fr/) and interviewee will deposit 2-3 times a month.

13:
Metadata considered important – where data obtained, quality of observations, purpose for which data collected. Technical and quality details.

14:
Table headings and technical details are added when table is ready.

15:
The team assigns the metadata.

16:
Interviewee makes data available on Astro-ph and own website. NASA-ADS and object specific databases will also add interviewee's work.

17:
Sharing of data is quite an accepted culture in astronomy. In addition it gets you known better and provides more collaboration opportunities.

The interviewee is happy to provide data on request as long as he/group is acknowledged.

18:
Would be reluctant to share data if it went unacknowledged.

19
No formal restrictions applied except the proprietary period restriction.

20:
Tables are on Open Access on the web. Interviewee thinks it better to give access to data than not, increases citations (see 17 above)

21:
Interviewee uses NASA-ADS. Also, for an undergraduate writing-up work last semester on some specific galaxies, Simbad was used to get to relevant publications.

22:
For teaching material, as above but also uses Google searches.

23:
Interviewee deposits in Astro-ph. Is not aware of any institutional repository [interviewer – Sheffield is part of the White Rose Consortium ePrints Repository White Rose http://eprints.whiterose.ac.uk/]

24:
Interviewee has a homepage with research links and goes from there to output repositories.

25:
Combination of simple and advanced searching.

26:
Sometimes needs to look at online help but tends to play around and has been using the repositories for years.

27:
Assistance is not requested of librarians or information professionals

28: Missing from source repositories?
In source archive, Hubble Space Data Archive is hosted in different places. If search on the US based webpages it is not as clear as ESO based webpages which has more options on which to search (area of sky, target, co-ordinates, instruments, wavelength restrictions)

29: Identification with perceived need of links from source to output?
Repositories want to see how cost effective they are and it's also useful to know if work you were planning has been done already.

30: Metadata difficulties?
Any source data available now is up to speed but older material may need metadata added. Quality control issues – going back might be difficult as the observational conditions not known.

31: Missing from output reps?
Interviewee would like full-text searching facility in NASA-ADS.

32: Benefit from ability to associate newly deposited publications with data from which derived?
Would be complementary to what's done already automatically but might be additional to as well because it's not happening everywhere yet.

33: New operations supported within an output repository – how meet your needs? Others advantageous?
Useful to know where data comes from and what was done with it. Quite happy with how things are working in the field at the moment.

34: dataset knowledgebase – value and specific issues
Probably useful but have to be careful with annotation if done by anyone. Takes quite detailed astronomer notes himself but gets less detailed from observatory.

35: Control necessary & access validation?
Source data - depends on if source data proprietary. No access control otherwise.

Also if get data from journals depends on whether a subscription is needed to get into the journals (interviewee mentioned older journal issues in particular).

US-Based Astronomers

Digital Data Preservation Workshop
JHU, Wed 26 July 2006

Sayeed Choudhury, Ethan Vishniac, Jan vandenBerg, Inga Kamp, Ray Lucas, Barbara Kern, Bob Milkey, Carol Christian, Alberto Conti, Imants Platais, John Grimes, Andrew Ptak, Dick Henry, Bob Hanisch

Phase in any policy that would require authors to provide digital data.

Metadata needed in addition to FITS keywords: subject, keywords, etc. Need to standardize metadata tags, like bandpasses. Would like to see processing history as part of metadata (e.g., how were images combined? How were images calibrated? Links to actual bandpass functions. Utilize VO characterization data model.

Will be multiple sites for data of different levels of processing.

Who supports first wave of authors in using such a system? Need astronomical expertise in journal editorial offices. Clifford Lynch uses the term “data scientist.” Working with first group of astronomers allows us to develop better tools and provide better documentation. Be pro-active with initial data contributors.

Most important scenario: have access to calibrated FITS file that is represented by an image or spectrum in a paper. Especially important for ground-based data, because space-based data is better archived.

Some think that it should be mandatory for authors to provide their digital data. In the long term, the data is the important thing; no one will care about the interpretation. Currently, different subdisciplines in astronomy have different expectations for data availability.

Controlled access (proprietary periods, embargos) can be necessary in order to protect data rights of authors.

Use small fraction of HST archival research funding to offset initial costs to authors?

Authors should give all reduced data back to the archives/data centers in addition to capturing the digital data represented in the journal.

Astronomers are pretty self-sufficient in finding information, especially using ADS. But, standardized keywords and controlled vocabulary would be very helpful. Find all images of galaxies published last year with $0.5 < z < 0.8$., for example. For finding papers, keywords are less important nowadays. However, the digital data needs to be discoverable and usable independently. Standard object classification nomenclature.

"Folksonomy" – community-based commentary and tagging. Wikipedia approach. Does this have a role here? Is astronomy community large enough? Taxonomy is time-dependent.

Data presentation and data inspection. "Integrity" issue. Presentation is important, but so is ability to inspect and allow reader to make an independent judgment.

Author threshold… Andy ok with uploading FITS files along with manuscript. Inga notes that many FITS files received by MAST have problems. Inga also agrees that it would be easy to ask for FITS file or ASCII file with spectrum. How much metadata will authors be willing to provide? Fill out form based on FITS header, and allow author to review and correct/augment. Build VO service that fills out data characterization metadata for a given dataset.

Journal readers: what's the minimum metadata to make the data useful? Coordinates, aperture, resolution, bandpass. Flux scale/calibration. Spatial resolution (PSF size). Noise characteristics (rms noise level). Spatial footprint (shape file). Date of observation. Exposure time. Characterize the data: single observation, combination, simulation, other derived product.

Simulations: input parameters, model outputs (density, temperature, fluxes, etc.). Simulation data itself is typically numeric table, with little or no metadata. Often just ASCII, or even a binary blob. Eventually, follow VO standards for theoretical data.

Community acceptance: everyone ready to take, no one wants to give! Will need incentives – funding, citations. For example, HST high-level products are used 10x more than standard pipeline products. Value-added; costs on the inputs side need to be kept low. Need tools to aid authors.

Distributed storage architecture: generally supported. Distributed approach includes the community in the process, provides robustness and reliability and multiple access points. Distributed systems should not be "hot tubs". Wikipedia works only because many people are reviewing the information.

Versioning is important if data or metadata changes or needs to be corrected.

Potential for re-use of data without attribution. Need for digital watermarks? Would need policy associated with re-use of data, part of agreement between author and publisher. Would be easiest if data is put in public domain or transferred to journal.