ISSN 1749-8767

DCC | Digital Curation Centre Case Studies and interviews

Wide field Astronomy Unit (WFAU)

http://www.dcc.ac.uk/resource/case-studies/wfau

Martin Donnelly
HATII, University of Glasgow,
Glasgow G12 8QQ
http://www.hatii.arts.gla.ac.uk

December 2005
Version 1.0

Legal Notices

Catalogue Entry

Title	DCC Case Study – Wide Field Astronomy Unit (WFAU)
Creator	Martin Donnelly (author)
Subject	Data curation; formats, processes and issues; interoperability; volume of data curated; system development; standards; legal factors; justification; Methodology, and Problems overcome; Human factors
Description	Case study on the Wide Field Astronomy Unit (WFAU), Edinburgh. Outlines data curation issues with which WFAU is involved, with an emphasis on interoperability. Particular regard is given to the transfer and reuse of data collected from disparate sources. The case study also covers other factors influencing data curation, including methodological development, standards and legal issues, evaluation, and human factors. A technical appendix outlines the technologies used in the development of the WFAU systems.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Joy Davidson and Andrew McHugh (editors)
Date	1 December 2005 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.2
Resource Identifier	ISSN 1749-8767
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines

Donnelly M, (December 2005), " Wide field Astronomy Unit (WFAU)", *Digital Curation Centre Case Studies and Interviews*, Joy Davidson and Andrew McHugh (editors), Retrieved <date>, from http://www.dcc.ac.uk/resource/case-studies/wfau

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC Case Study and Interviews

Editors

Joy Davidson
Humanities Advanced Technology and Information Institute, University of Glasgow (UK)

Andrew McHugh
Humanities Advanced Technology and Information Institute, University of Glasgow (UK)

Peer Review Board

Case Study: Wide Field Astronomy Unit, University of Edinburgh (http://www.roe.ac.uk/ifa/wfau/)

This Digital Curation Centre case study is the result of an email questionnaire completed by Dr Bob Mann and Dr Peredur Williams of the Wide Field Astronomy Unit in June/July 2005, and on subsequent email communication between Dr Mann and Martin Donnelly of the DCC and HATII, University of Glasgow.

Table of Contents

Executive Summary

The Wide Field Astronomy Unit (WFAU) was formed in 1999 as part of the restructuring of the Royal Observatories. Projects in which WFAU is involved include the development of the WFCAM Science Archive, the VISTA Science Archive, and AstroGrid - the UK's contribution towards building an international 'Virtual Observatory.'

Until recently, astronomers had to inspect photographic plates, films and catalogues in person. In the 1970s a programme of scanning plates began, providing digitised versions of sky surveys. Users are now able to download images of portions of scanned plates and catalogues from the Web. In addition to creating and curating these data, WFAU serves a large community of data re-users. By linking its databases with the nascent Virtual Observatory, WFAU is making its data available to the world's 10,000 astronomers. Regular curation tasks include loading catalogues into a database, matching them with prior observations, and preparing data for publication via a web interface. Data also occasionally need to be recalibrated, or replaced when superseded through new observations.

The data products extracted from the WFAU archives are held in standard formats, primarily the Flexible Image Transport System (FITS) format. FITS is not overly prescriptive in terms of what is recorded in the metadata headers. There is no set schema, hence each data centre is able to decide what it should record. The desire for wide applicability has left the set of FITS keywords which can be used in the metadata records defined only by weak constraints on which, and how many, characters they can contain, rather than being selected from a controlled list. As a result, the metadata records in FITS files can be difficult to interpret beyond the institution or collaboration which generated them. This terminological looseness can create potential problems for data centres which ingest quantities of externally created data.

Significant effort has been put into the curation of data beyond simple preservation. A great advantage that data curation processes have for adding value to pre-existing data is in pooling/aggregating information from different sources: it is now possible for users to match up objects across a number of large sky survey datasets. The investment that the astronomical community is putting into the development of the Virtual Observatory illustrates the strength of the continuing belief that the curation of digital data is important to the future of astronomy. Similarly, the advent of e-Science is a reflection of digital data curation's importance to science more widely, given that the 'data avalanche' experienced by many scientific disciplines has driven the UK e-Science programme. Researchers across a wide range of domains are, like WFAU, finding themselves faced with new and challenging responsibilities for curating digital data.

Introduction

The Wide Field Astronomy Unit (WFAU) is part of the Institute for Astronomy (IfA), one of the research groups within the School of Physics in the University of Edinburgh, Scotland. The Unit was formed in 1999 with the transfer to the IfA of the existing sky survey group within the Royal Observatory Edinburgh (ROE), as part of the restructuring of the two Royal Observatories. The Institute is housed in the ROE site, which it shares with the UK Astronomical Technology Centre (ATC) and the ROE Visitor Centre.

The bulk of the funding for WFAU's staff, which currently numbers thirteen, comes from the Particle Physics and Astronomy Research Council (PPARC). WFAU has a rolling grant from PPARC, renewed at three-yearly intervals, and additional grants exist for ongoing projects, such as the development of the WFCAM Science Archive (WSA), to hold data from the Wide Field Camera – WFCAM - on the UK Infrared Telescope in Hawaii), the VISTA Science Archive (VSA), and work on AstroGrid. AstroGrid is the UK's contribution towards building an international "Virtual Observatory" (VO), which will be a federation of astronomical data and computer resources worldwide. AstroGrid is a partner in the four-year EU-funded VOTECH project, which is undertaking R&D work towards the development of a Euro-VO, and part of this work is being undertaken within WFAU. In the Financial Year 2004/05, WFAU received funding of approximately £700,000, around 90% of which came from PPARC and the remaining 10% from the EU. About 80% of this funding supports sky survey archive development and operations, while the remainder covers development work on the Virtual Observatory.

Data curation

The sky survey group in Edinburgh has been curating data and serving them to the astronomical community for more than thirty years. Most of this work has been done in connection with the UK Schmidt Telescope[1] (UKST) in Australia, which was run directly from ROE until 1988, and thereafter by the Anglo-Australian Observatory in Sydney in collaboration with ROE. Between 1973 and 2002, UKST conducted major systematic photographic surveys of the southern celestial hemisphere, as well as conducting many smaller observing programmes for individual astronomers. Together the surveys and other programmes have yielded more than 19,000 original photographic plates and films, which are now stored in the ROE's Plate Library, along with copies of other sky atlases.

Initially, astronomers had to inspect these photographic plates visually, but in the early 1970s the ROE began a programme of scanning plates with specially developed automatic *densitometers*, first GALAXY, then COSMOS, and, from 1992, SuperCOSMOS, which greatly surpasses its predecessors in terms of scanning speed and precision.[2] These machines were designed primarily to provide digitised versions of the systematic sky surveys, and the scanning programme has culminated in the SuperCOSMOS Sky Surveys (SSS). These provide astronomers with access to digitised sky survey data in B (blue), R (red) and I (near-infrared) passbands and from the Hα spectral line. Users can download images of portions of scanned plates and/or catalogues listing attributes (size, location, shape, brightness, etc) characterising celestial objects detected in the image data by an automated image analysis program.

Like the image analysers used with previous astronomical sky surveys, this program searches for pixels significantly brighter than the background sky and connects them into objects, which are then characterised by a range of attributes derived by fitting an elliptical model to the pixel data. 'Deblending' software seeks to separate objects which overlap in the pixel data, and post-processing

[1] Schmidt telescopes are designed to have wide fields of view, making them ideal for sky surveys.

[2] A densitometer is an instrument for measuring the darkness, or light-stopping power, of exposed and processed photographic material - i.e. its density. Definition: http://www.screensound.gov.au/glossary.nsf/Pages/Densitometer

of the object catalogues includes classification, distinguishing stars from galaxies. The surveys

contain data for the southern celestial hemisphere in B, R (at two epochs[3]), and I, while, for the northern hemisphere, SuperCOSMOS has scanned B and (two epochs of) R plates from the Palomar Observatory Sky Surveys and is close to completion of its I-band survey. When complete, in 2006, the SSS will comprise a multi-epoch, multi-colour survey of the whole sky.

Data curated: formats, processes and issues

The primary focus of WFAU's work has recently shifted from traditional photographic surveys to the new generation of sky surveys produced by digital detectors. Digital detectors – charge-coupled devices for the visible portion of the spectrum, plus their analogues for the near-infrared[4] – have been used in astronomy for a long time, but only recently has it been possible to make them large enough to survey significant areas of sky, so born-digital sky survey data is a relatively new development.

WFAU is now responsible for curating the data from two large near-infrared survey instruments: the WFCAM in Hawaii, and the VISTA telescope being built in Chile. Most of the WFCAM's operation time is dedicated to the UK Infrared Deep Sky Survey (UKIDSS), which started taking data in early 2005, and which is expected to continue for roughly seven years, while the VISTA surveys are planned to run for approximately twelve years from 2006/07 to 2018/19. From the telescope in Hawaii, data are transferred to a processing

centre in Cambridge, where instrumental signatures are removed, to yield clean images of the sky, and catalogues of sources extracted from them. WFAU's regular curation tasks include loading these catalogues into a database, matching them up with observations of the same objects taken previously by the WFCAM and from other sky surveys, and preparing the data for publication via a web interface in a manner that suits the users of the WSA. Further curation tasks include the occasional recalibration of data in the database, and the replacement of data when they are superseded through new observations.

The bulk of the digital data currently curated by WFAU comes from the SuperCOSMOS scanning programme, which yields a set of "housekeeping files" recording metadata for each plate scan, currently being incorporated into the SuperCOSMOS Science Archive (SSA). The metadata capture for the plate-scanning programme with SuperCOSMOS is mostly automated, with some manual interventions. The scanning procedure – and subsequent processing of the scanned images to yield catalogues of detected sources – is controlled by a series of computer programs, supervised by a human operator who inputs certain parameters (e.g. plate number). These are recorded in the 'housekeeping data,' together with a much larger quantity of metadata generated automatically by the programs as they run.

The chief issue in the re-use of archival astronomical data is whether a certain feature is real (in the sense of reflecting something in the sky when the observation was made), or a blip (or 'artefact') produced by the hardware or software during the processing that takes place between the light from the sky hitting the detector and the astronomer extracting the data product from the archive. Data processing attempts to identify known classes of artefact, but this knowledge is imperfect (although it

[3] The epoch of an observation is the time at which it was made. Not all plates within each of the constituent subsurveys of the SSS (e.g. the B or I survey) were observed at exactly the same epoch, but the temporal baseline between plates in the same field on the sky from different subsurveys is sufficient for the study of time-dependent phenomena.

[4] A charge-coupled device (CCD) is a light-sensitive integrated circuit that stores and displays the data for an image in such a way that each pixel (picture element) in the image is converted into an electrical charge the intensity of which is related to a colour in the colour spectrum. Definition: http://searchstorage.techtarget.com/sDefinition/0,290660,sid5_gci295633,00.html

improves with time as experience of using particular instruments grows.) It is therefore important to record the versions of all data processing software packages used in the generation of a particular data product. (For example, it may be that a certain artefact was

only successfully removed in the software version released on a certain date, or that a particular version of a software module had a bug in it which was fixed at a later date.) Ideally, a data centre would only publish data products with all known problems removed, but in practice this is not always the case, and without a full record of the data's history it can be impossible to identify the cause of the problem, or to identify which data were affected by it.

Apart from these issues related to data processing, similar problems can arise from the original observations – e.g. instrumental settings in particular periods being non-optimal, or data taken under certain observing conditions being known to suffer from particular defects – so the re-user of the archival data really wants to be able to trace the history of a data product all the way back to the telescope, in order to have confidence in its properties. This allows users to 'drill down' from a particular attribute value and trace its provenance from the exposure of a particular glass plate (on a night with recorded conditions), through a data processing pipeline (with known parameter settings).

Data are returned from both the SSA and the old SSS data-access systems in standard (non-proprietary) astronomical formats, which include inbuilt metadata. The data covered by the WFAU's preservation/curation strategy comprise image and tabular data sets. Within the data centre, much of the data is held in proprietary formats (e.g. a commercial RDBMS or home-grown binary data formats), but in all cases the data products extracted from the archives are in standard formats; primarily Flexible Image Transport System (FITS) format, although smaller tabular datasets can also be extracted in VOTable, an XML data format developed within the Virtual Observatory community as an alternative to FITS.[5]

FITS is the principal standard format used in astronomy, and as the name suggests it was originally designed to aid the transfer of image data sets between observatories. The FITS standard has since been developed to handle other types of data, notably tabular datasets, either in ASCII or in plain binary formats. VOTable retains some of the structure of FITS files for tabular data, with a metadata section included alongside the table of data values, meaning files can be self-documenting to a fair degree. However, in their desire to make a simple format, the VOTable designers produced slightly unusual XML (reminiscent of a table in HTML), and like all XML it is too verbose to store the large datasets common in astronomy.

The FITS standard is not overly prescriptive in terms of what is recorded in the metadata headers which form part of every data file. There is no set schema, hence each data centre is able to decide what it should record. To enable confident reuse of data, SSS metadata includes every attribute/parameter that could have a bearing on the data quality, and which a user might want to know. Therefore the images and the source catalogues exported from the SSS all contain long FITS headers recording metadata about the observation and reduction process. The same strategy was followed when the source catalogues started being ingested into a relational database to form the SSA, so there are tables in the SSA which store all available metadata relating to each SuperCOSMOS plate scan. Data products exported from the SSA are not automatically accompanied by those metadata, although the user can opt to receive these.

[5] http://www.ivoa.net/Documents/latest/VOT.html Both VOTable and FITS are open formats: their full specifications are freely available on a royalty-free basis, and those specifications were developed (and are maintained) through a community process, open to all interested parties.

In terms of data formats, the feeling is that, so long as all of the data exported to users is in standard formats (which are almost all non-proprietary), it doesn't matter what is used to store the data within the data centre. As Dr Bob Mann of WFAU says: "If we judge that we give our users a better service by ingesting the data from FITS files into a commercial database than by keeping them on disk as FITS files, then that's what we'll do - so long as the data are exported from the database into a standard form for delivery to the user."

Interoperability

This desire for wide applicability has left the set of FITS keywords which can be used in the metadata records defined only by weak constraints on which, and how many, characters they can contain, rather than being selected from a controlled list. While some keywords relating to certain particularly important concepts (e.g. the area of the sky covered by an image) have developed into *de facto* standards, a variety of keywords can be used to represent the same quantity. As a result, the metadata records in FITS files can be difficult to interpret fully beyond the institution or collaboration which generated them. This terminological looseness can create potential problems for data centres – such as WFAU – which ingest quantities of externally created data.[6] For example, WFAU's basic WFCAM and VISTA data products are generated by a data reduction pipeline run in Cambridge, and then copied to Edinburgh for loading into the archives through which users will access them. The flexibility of FITS keywords tends to be seen as an advantage in a field where much of the metadata to be recorded is specific to a particular instrument or sub-area of astronomy. It would be difficult to develop from scratch a fixed list of FITS keywords which would meet the requirements of all users. Within the Virtual Observatory movement, however, there are plans to introduce some sort of semantic framework against which values recorded in database or other digital files can be mapped. This is taking place in two ways. Firstly, there is a set of quantities called Unified Content Descriptors (UCDs), which comprise a controlled vocabulary of terms covering all physical quantities recorded in astronomical data sources. Their aim is to aid interoperability between data sources, likely to have been developed in isolation. It is therefore likely that values of a particular physical quantity (e.g. the brightness of stars in a particular passband) are recorded in two separate data sources, but with different names. This anomaly need not hinder their inter-comparison, provided the appropriate column in each database is tagged with the same UCD, acting as a sort of semantic type for that column, and indicating what can legitimately be done with those data values.

The UCDs are a bottom-up approach to adding semantic content to aid interoperability and re-use of data within the Virtual Observatory, but there is also a top-down approach being undertaken by IVOA's Data Modelling working group. The aim is to develop a comprehensive data model for all the concepts relevant to astronomy. It is likely that this will eventually be expressed as an ontology, although, in Dr Mann's words, astronomers tend to be scared by "the O-word."

To facilitate the transfer of data between Edinburgh and Cambridge, an Interface Control Document (ICD) was developed by staff to detail the FITS metadata keywords used in data files transferred between the two sites, allowing WFAU to safely run the curation

[6] One of the main justifications for specialist data centres like WFAU is that they have particular expertise in the kinds of data that they curate, and are therefore able to support users manipulating such data. If the metadata recorded in products curated by a data centre are non-standard, the ability of centre staff to provide the necessary support to subsequent data re-users may be compromised. This difficulty is exacerbated over time, as the availability of knowledge with regard to the meaning of particular project-specific metadata is likely to diminish. However, the very reason why a flexible (and, therefore, loose) vocabulary is used for FITS keywords is the perceived difficulty of defining a data model which includes all the quantities that may need to be recorded in the metadata for data products from current - let alone future – astronomical instruments. The best practical solution may be to include (and preserve) sufficient documentation alongside each use of a non-standard FITS keyword within a header.

tasks specified by the metadata records in each FITS file. All data would be transferred between Cambridge and Edinburgh in the form of Multi-Extension FITS (MEF) format, a variant of the FITS standard in which several datasets are included in the same file, which is well-suited to the way in which data emerge from the WFCAM, which comprises four detectors. The ICD specifies the FITS keywords to be used in each header (i.e. metadata) section, the physical units to be used in each image section, and the naming convention for the MEFs themselves.

Volume of data curated

The SSS catalogue data were originally stored in flat files on disk, but are now being ingested into a relational database management system (RDBMS), to create the SuperCOSMOS Science Archive (SSA). WFAU currently manages approximately 7TB of data in the relational database, 5TB of data products in flat files on disk, and in excess of 50TB of flat files on tape (including multiple back-up copies of the RDBMS and data products.) The advent of WFCAM and VISTA will significantly increase the volume of WFAU's data holdings. WFCAM will generate 10-20TB of image data per year (capable of lossless compression by approximately a factor of three for storage), which will yield 1-2TB of catalogue data annually for at least the next seven years.[7] VISTA will have a data rate roughly double that of WFCAM. By contrast, if all the plates in the ROE Plate Library – the results of well over thirty years of observations – were to be scanned by SuperCOSMOS they would yield only around 40TB of image data.

System development

The team developing WFAU's systems generally have backgrounds in astronomy, rather than IT. Some retain a time allocation for personal research, while others undertake data-centre work full-time. The short timescales involved in the production of the SSA and WSA dictated that their development be undertaken by astronomers, rather than IT professionals who would have to learn the science drivers and usage modes for the archives on the job. However, as the focus shifts from the development of the basic science archive system to its enhancement for handling the increase in data volume from WFCAM to VISTA, the mix of skills needed in the development team also shifts.

The design and development of the SSA and WSA benefited significantly from the advice of the team who developed the archive for the Sloan Digital Sky Survey (SDSS), the first of the new generation of born-digital sky surveys. The SDSS archive team were therefore the first to solve many of the problems faced by

the SSA and WSA, such as how to implement within a commercial RDBMS a scheme for efficiently indexing data distributed on the surface of a sphere. The SDSS archive team is based at the Johns Hopkins University, Baltimore, USA, and led by Professor Alex Szalay.[8] They in turn have benefited from an ongoing collaboration with Jim Gray of Microsoft Research since 2000.[9] Dr Gray has also provided invaluable (and unpaid) advice to WFAU, working closely with staff in the initial design of the SSA and its implementation using Microsoft's SQLServer RDBMS.

Standards and Legal factors

In addition to creating and curating data, WFAU also serves a large community of data re-users. There are around 10,000 professional astronomers worldwide, and by hooking their databases up into the nascent Virtual Observatory, WFAU is making these data available to the whole population. Most astronomical data are eventually placed in the public domain, but it is typical for there to be a proprietary period of one or two years after

observations are made within which data access is restricted to the individual astronomer or collaboration who made the observations, or to the community served by the funding agency that paid for them. WFAU

[7] One terabyte of data is roughly equivalent to 250 million pages of text.

[8] http://tarkus.pha.jhu.edu/~szalay/
[9] http://research.microsoft.com/~Gray/

therefore has to implement a range of proprietary rights. No proprietary rights exist over the SSS data, so the SSA is open to anyone to use. The bulk of the data that will enter the WSA comes from the UK Infrared Deep Space Survey, comprising astronomers from the member states of the European Southern Observatory (ESO) and Japan. WFAU is currently implementing an authorisation scheme whereby access is restricted to registered users from the UKIDSS community. WFCAM time is also awarded[10] for non-UKIDSS programmes, and so WFAU must also restrict access to the data from each of these to its Principal Investigator until the proprietary period ends.

WFAU's operations to date have not been significantly influenced by legal/copyright/intellectual property/data protection/freedom of information factors. Astronomical data has essentially no commercial value; the surveys whose data are curated by WFAU are conducted as a service to the astronomical community, and many areas of astronomy rely on the ready sharing of complementary datasets, so - in common with other astronomical data centres - they have taken a relaxed attitude to these issues to date. However, the WFAU realises that recent developments, such as the new database right, may require more careful and more explicit handling of these matters.[11]

Justification

Astronomical data have legacy scientific value: while a new detector may take a 'better' image of a region of the sky, some celestial phenomena are transient and others time-dependent, hence their study requires data from multiple epochs of observation which can span long periods of time. So while the photographic sky survey plates are decades old – and tend to produce data of inferior quality to modern, digital detectors – they have lasting scientific value. The guiding principle is to keep these legacy data accessible. This has involved migrating the data between several generations of media. (For example, the original SuperCOSMOS image files are currently stored on exabyte tape, and the WFAU is about to embark on a process of transferring them to LTO-3 tapes, before the ability to read exabytes safely at the ROE is lost.)[12]

Significant effort has been put into the curation of these data beyond simple preservation. Work has centred on updating the ways in which users access the data, and integrating them with newer datasets. In both cases this work is designed to facilitate the continued scientific exploitation of the data. Perhaps the greatest advantage that data curation processes have for adding value to pre-existing data is in pooling/aggregating information about particular celestial objects from different sources. Observations made of an object in different regions of the spectrum often probe different physical processes, thereby providing complementary information to the astronomer, who will often want access to all the data that exist on a particular object or set of objects. Facilitating such studies is one of the key aims of the Virtual Observatory, but some of this work is undertaken within single data centres. It is now possible for users to match up objects in the SSA with those in a number of other large sky survey datasets, making use of local copies of those surveys.

Updating the means of user access to the data has involved the ingestion of SSS data into the SSA, providing users with a much wider range of query functionality. In the old SSS system, users could extract a catalogue

[10] Time on public telescopes is awarded competitively: astronomers write proposals, justifying their request for a certain amount of observing time on the basis of the science they intend to do with the data they will obtain.

[11] The *database right* acts in tandem with *copyright*. Databases are subject to database right when it can be demonstrated that the contents of a database are the product of significant investment. See http://www.intellectual-property.gov.uk/std/resources/other_ip_rights/database_right.htm

[12] The majority of WFAU's 19,000 photographic plates and films have not been scanned into digital form, and care is taken to keep them in good condition so that the information they store in analogue form is not lost. Other non-digital holdings include the archives of the UK Schmidt Telescope, which are held on a mixture of media and are primarily administrative in nature, although a small proportion may also have heritage value.

of objects from a given area of sky, but were unable to query the whole SSS dataset for objects with a particular set of properties; this can be done straightforwardly with the SSA. Integration of the SSS data with other sky survey datasets involves preparing additional database tables in the SSA, through running a spatial matching code. This allows users to link entries between the SSA and other popular sky survey databases, thereby enabling them to merge data for a single object from two separate databases, increasing efficiency. The SSA is also one of the first sky survey databases to be linked to the nascent Virtual Observatory, ensuring that SSS data will remain available as VO data access methods and analysis tools develop.[13]

Methodology, and Problems overcome

WFAU's preservation methodology to date has been quite simple, in that it has been possible to preserve *all* the data generated by the SuperCOSMOS plate-scanning programme; they aim to continue this for the born-digital WFCAM and VISTA surveys. This policy is driven by the requirement that data products can be re-used by other users in the longer term, in many cases for purposes very different from that for which they were originally obtained. For example, one part of the SDSS was a spectroscopic survey to determine distance to ~10^6 galaxies, so that the strength of their spatial clustering could be measured in 3D-space, rather than just in projection on the sky, and the spectra thus obtained have subsequently been analysed to reveal the evolutionary history of the galaxies.

One constraint faced early in the programme was financial: the cost of spinning disk was initially too high to store all the SuperCOSMOS image data in uncompressed form. It was therefore decided that the image data served to users should be in a 20-times compressed format. This was generated using a lossy compression algorithm, but much of this "lost" information is actually noise in the brightness level of the background sky, so compression does not significantly degrade the image data for scientific purposes – in any case, the object catalogue data are generated from the uncompressed image data. Since then, good lossless image compression methods have been developed for astronomical image data, and these will be applied to WFCAM and VISTA image data. The algorithm used by WFCAM – Rice compression – will compress images by a factor of about three, which is now sufficient given that disk space is now much cheaper than in the past.

Human factors

A significant issue for WFAU has been the steep learning curve faced by the staff developing the SSA and WSA. Whilst WFAU have been curating digital data for more than a decade, the scale of problems presented by the WSA and VSA, and the technologies required to solve them were entirely new; according to Dr Mann, learning about relational databases by building one that is over a terabyte in size is something of a baptism of fire. Some preparatory training would have been beneficial (and, indeed, some – in relational database design – was obtained, through the National e-Science Centre), but would probably not have reduced the steepness of the learning curve faced by WFAU's staff significantly, since the challenges in applying standard technologies to a particular scientific domain, such as astronomy, differ in many respects from those in commercial domains on which training examples are likely to be based.

WFAU has recently recruited a dedicated science archive curator, who does not have an astronomical background. It is expected that the Unit will continue to require staff from a range of backgrounds, and with a range of skills: the technical requirements of WFAU's science archive curation greatly exceeds that which can be comfortably provided by professional astronomers.[14]

[13] The International Virtual Observatory Alliance: http://www.ivoa.net/

[14] Experience with software developers hired from the IT industry to work on AstroGrid suggests that the appointment of staff without an astronomical background is not problematic. Someone who had studied astronomy as an undergraduate would have no more knowledge of practical questions (such as how the data used in astronomical research are obtained, stored and

Evaluation

Qualitative evaluation of work to date has taken place as part of WFAU's grant renewal process. Plans for the extension/expansion of the WSA system in readiness for VISTA will be formally reviewed in Spring 2006, and these will include evaluation of the WSA. The success of the work will only really be revealed over the coming years, as the WSA starts to be used in earnest. However, the fact that there was a robust system ready to ingest the first WFCAM data as it came off the telescope is a success in itself, given the speed with which WFAU staff have had to get to grips with a range of new technologies. The SuperCOSMOS scanning programme is now nearing completion and will be rounded off by the securing of the digital plate scan data through their migration to LTO-3 tapes. WFAU's current plans focus on further development of the science archive system prototyped with the SSA, and extended for the WSA, in preparation for VISTA.

WFAU is also looking ahead to possible involvement in future sky surveys, and to enhancing the services it provides its user community, especially in facilitating the analysis of data through the Virtual Observatory. Many of the analyses that astronomers want to run on sky survey databases require more data than they can readily download to their local workstations, so it would make more sense for users to be able to upload data analysis code to be run at the data centre. Clearly, though, this has many possible security risks, so it would be very useful for there to be some work on strategies by which data centres could offer users flexibility in the analysis of their databases without risking their integrity. This seems like one area where the role of the curator is expanding, and research in this area would be welcomed by the WFAU.

Summing up

The investment that the astronomical community is putting into the development of the Virtual Observatory illustrates the strength of the continuing belief that the curation of digital data is important to the future of astronomy, since the heart of the VO is the series of digital data resources that it federates. Similarly, the advent of e-Science is a reflection of digital data curation's importance to science more widely, given that the "data avalanche" experienced by many scientific disciplines has driven the UK e-Science programme. Researchers across a wide range of domains are, like WFAU, finding themselves faced with new and challenging responsibilities for curating digital data.

manipulated) than someone hired from industry, who would be likely to have experience of similar issues in other application areas. In short, it is easier to teach astronomy to an IT expert, than IT to an astronomer.

Annex I – Technological issues

WFAU employs a range of technologies, largely as a result of the rapid evolution in data curation responsibilities and the necessity of learning on the job by prototyping solutions, rather than being able to implement an existing design. The original SSS data (image and catalogues) are stored as flat files in bespoke formats on a RAID-5 array and are accessed by users querying Web forms.[15] The RAID array is attached to an HP Alphaserver running Tru64 Unix. The design for this system was developed in collaboration with a hardware supplier in preparation for a funding bid several years ago, based on current thinking at that time.[16] The SSA and WSA are implemented in SQLServer on dual-processor Xeon servers connected to Ultra320SCSI disk arrays via 4-channel hardware RAID controllers. Both systems employ spanned RAID – i.e. there are individual RAID5 arrays on each SCSI channel, and then a RAID0 stripe across these to produce what is sometimes called a RAID50 spanned array. This design was developed largely in-house as the result of a series of tests designed to produce a robust system with a high I/O performance. The images from which the WSA catalogue data were extracted are stored as FITS files on a RAID-5 array of SATA disks connected to a Linux server. An LTO-2 drive is used to generate tape back-ups of all catalogue data from the SSA and WSA data.

Hardware advice came from a local vendor, Eclipse Computing, with whom WFAU have worked for a number of years, and from Andy Knox of IBM, who is seconded half-time to the National e-Science Centre. Both were invaluable in helping WFAU to plan and run the tests which influenced the WSA hardware design.

[15] RAID (redundant array of independent disks) is a way of storing multiple copies of same data in different places across multiple hard disks. Definition adapted from: http://searchstorage.techtarget.com/sDefinition/0,290660,sid5_gci214332,00.html

[16] If the project were starting from scratch today, the SSS catalogue data would not be stored as flat files on disk. Instead, a relational database would be used, as for the SSA, although the images would be kept as flat files on disk. Then, instead of having a single server for handling both catalogue queries and image extraction requests, the hardware would be decoupled into a database server and an image server, as for the WSA.

Annex II – Abbreviations

Abbreviation	***Full phrase***
ATC	Astronomical Technology Centre
FITS	Flexible Image Transport System
ICD	Interface Control Document
IfA	Institute for Astronomy
MEF	Multi-Extension FITS
PPARC	Particle Physics and Astronomy Research Council
ROE	Royal Observatory Edinburgh
SDSS	Sloan Digital Sky Survey
SSA	SuperCOSMOS Science Archive
SSS	SuperCOSMOS Sky Surveys
UCD	Unified Content Descriptor
UKIDSS	UK Infrared Deep Sky Survey
UKST	UK Schmidt Telescope
VO	Virtual Observatory
VSA	Vista Science Archive
WFAU	Wide Field Astronomy Unit
WFCAM	Wide Field Camera
WSA	WFCAM Science Archive