Large Scientific Data Systems:

analysis of some existing projects and their applicability to Gaia

Treball presentat per William O'Mullane

William.OMullane@sciops.esa.int

European Space Astronomy Centre, Villafranca del Castillo, Madrid

i dirigit pel **Dr. Xavier Luri Carrascoso** per optar a la obtenció del DEA

*Departament d'Astronomia i Meteorologia, Universitat de Barcelona***.**

Table of Contents

1 Introduction

This study concerns existing astronomical systems and their relevance to the Gaia mission. The author has recently taken up the post as Gaia Science Data Processing Manager and has undertaken a survey of some existing science projects to assist in the final decision on the Gaia core system design as well as the management of the Gaia Data Processing and Analysis Consortium (DPAC). Many ideas are presented here, backed up by content from the survey.

One problem with a technology study is of course the rate of change of technology and the possible staleness of some of the solutions. Astronomical projects tend to span decades – the technology selected at the beginning of a project may not be the technology of choice to carry out such a project again at the end. Within the lifecycle of the project itself this problem of staleness needs to be dealt with, how may nimbleness be included in such a large undertaking? Work practices, management of people and software may be transferable. There may be lessons to be learned.

A brief overview of Gaia its instruments and the complexity of the processing are provided in Section 2.

The author conducted a series of interviews with key individuals in many projects and these are summarised in Section 3 and presented in full as appendices.

Some thoughts on applicability to Gaia are presented in Section 4 and conclusions are draw in Section 5.

2 Gaia

Gaia is an incredibly exciting cornerstone mission of the European Space Agency (ESA). ESA is due to launch the 2000kg Gaia satellite in 2011 on a Soyuz-Fregat rocket. It consists of two Astrometric instruments as well as a photometric-radial velocity instrument allowing it to build a phase space map of our galaxy. One may trivialise Gaia saying it is simply Hipparcos II, yet it is so much more. Hipparcos accurately observed 140,000 objects whereas Gaia will observe closer to one thousand million galactic and extra-galactic objects. The accuracy predicted for Gaia is also unprecedented, in the microarcsecond range, it will observe fainter objects than Hipparcos down to down to G=20 magnitude (where G is the passband of the astrometric instrument). The addition of the radial velocity instrument addresses a major shortcoming of the Hipparcos mission allowing correct velocities in all three dimensions to be calculated. The potential scientific benefits of Gaia are practically innumerable. The data processing required to produce a Gaia Catalogue from which these scientific benefits may be reaped is, however, non trivial.

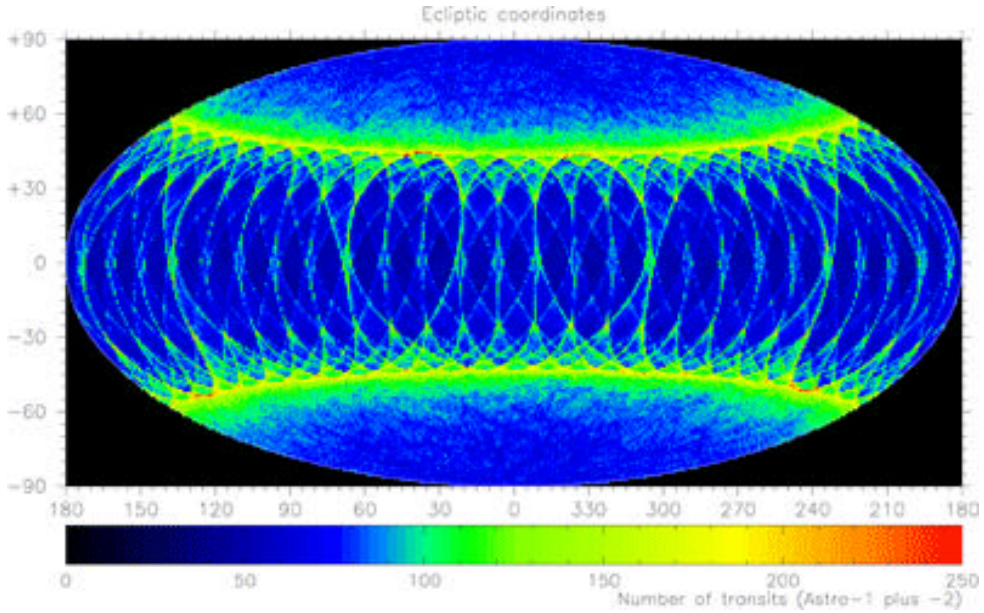

Figure 1. Predicted number of transits over the five-year mission (Jos de Bruijne/ESA)

2.1 The Mission

Gaia will be injected into a Lissajous orbit around the Sun-Earth Lagrange point L2, where it shall spin for five years observing the whole sky, conducting a census of one thousand million objects, observing each approximately one hundred times. Figure 1 plots the predicted sky coverage for Gaia's Astrometric telescopes showing a minimum of fifty transits over most of the sky. A transit for Gaia means an object crossing the focal plane.

2.2 Astrometric Instruments

The satellite contains two Astrometric telescopes [26] with a fixed angle of between 90 and 106 degrees between them, the exact angle must still be finalised. The viewing directions of both telescopes overlap on a common focal plane. The entrance pupil is 1.4 m x 0.5 m and the focal length is 46.67 m for each Telescope. Figure 2 shows the two primary Astro mirrors and the focal plane they eventually reflect onto.

The Astro focal plane [27] functionally consists of three CCD strips: the Astrometric Star mapper (ASM), the Astrometric field (AF) and the Broad Band Photometer (BBP). The mosaic contains 180 Charge Coupled Devices (CCDs) with pixels of 10 micrometers along scan x 30 micrometers across scan size (44.2 mas x 132.6 mas). The first two columns of CCDs form the ASM, which works out transit of objects crossing the focal plane, thus allowing efficient read-outs of the CCDs in the main focal plane. The main Astrometric measurements are made in the AF in the next eleven columns of CCDs. The final five columns of CCDs form the BBP, which provides multi-coloured photometric measurements for each object. Figure 3 depicts the Astro focal plane.

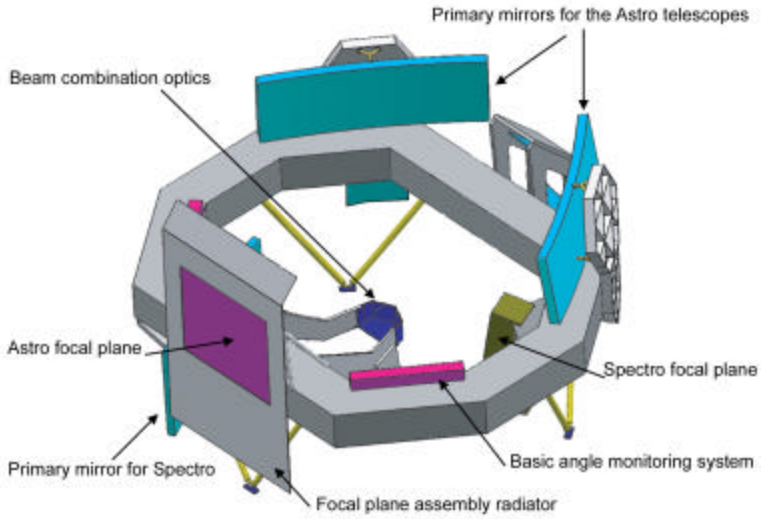

Figure 2. The scientific payload. (Astrium)

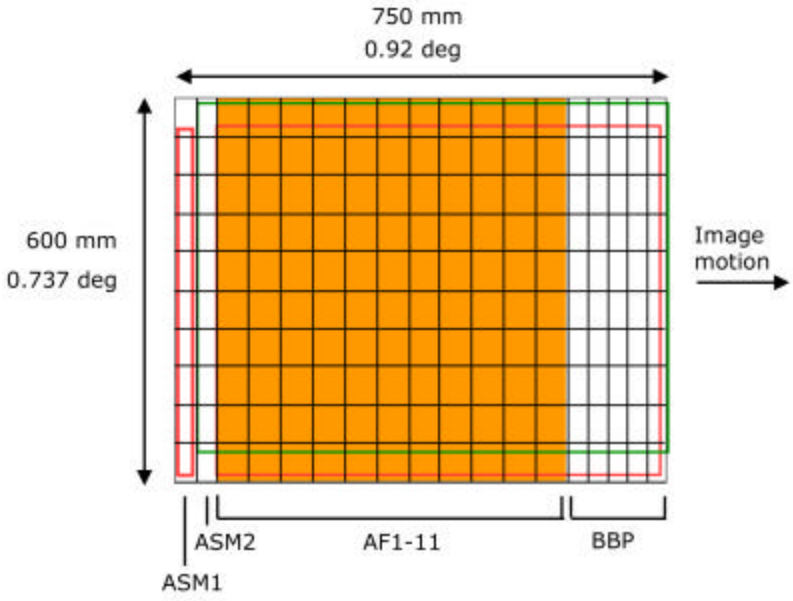

Figure 3. The Astrometric focal plane showing the 18 strips of CCDs and indicating the direction in which objects will cross the plane. The large red box is the field of the Astro-1 telescope while the large green box is that of Astro-2. (Astrium)

The predicted accuracy for Gaia astrometry is a few microarcseconds to G~12. From G=12 to G=20 noise will have an effect but 20-25 microarcseconds are expected to G=15 ranging to a few hundred microarcseconds at G=20. These accuracies are only possible given the statistical effect of having ~100 measurements per object, the focal plane itself will not make such an accurate individual measurement.

2.3 Photometric and radial velocity instruments

Gaia contains a third telescope which feeds the so called Spectro instrument, composed itself of two instruments. The first instrument is the Radial Velocity Spectrometer (RVS) primarily designed for the acquisition of radial velocities [30] and the second is the Medium Band Photometer (MBP). The telescope viewing direction is in the same plane as the astrometric instruments but at an angle of 38 degrees from Astro1. The entrance pupil is 0.25m^2 and the focal length is 2.1m [32]. The Spectro primary mirror and focal plane positions are depicted in Figure 2.

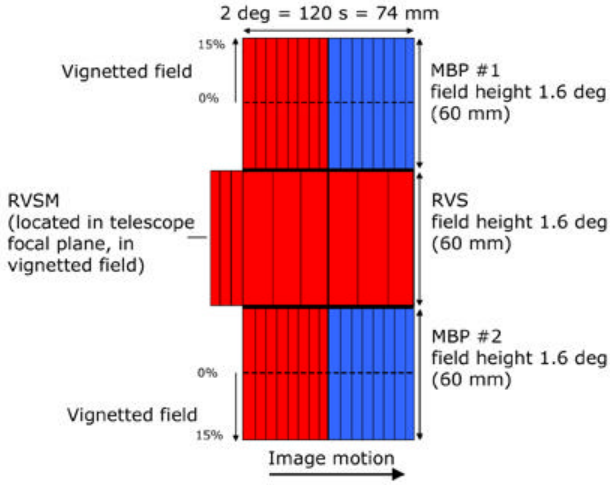

Figure 4. Spectro focal plane showing the spectrograph and the two Medium Band Photometers (MBP). The RVS sky mappers (the block on the diagram is actually smaller than it should be) precede the spectrograph which is aligned with the MBPs[1]. (Astrium)

[1] The focal plane depicted in this figure is now obsolete a new figure was not available at time of submission.

The Spectro focal plane (Figure 4) is in two distinct physical planes. The RVS Sky-mappers and photometric detectors are located at the telescope focus while the spectroscopic instrument CCDs are located at the spectrograph focus [31]. The RVS Sky-Mappers consist of 8 CCDs of 336x3930 pixels. As with the astrometric instruments the RVS Sky-Mappers are used to select which pixels need to be read out from the spectrograph field of view. The RVS Sky-Mappers provide Magnitudes in the "spectrograph photometric band" and are also used to detect Near Earth Objects.

Each of the MBP blocks consist of 2 sets of 8 CCDs (similar to the Sky-Mapper CCDs). Again the first CCD or two (still TDB) will be used as Sky-Mappers and the remaining CCDs will be coated with 11 different Photometric filters.

The Spectrograph focal plane consists of 6 Low Light Level CCDs of 1010x3930 pixels.

RVS will provide radial velocities and about 100 individual spectra for up to 250,000 million stars. This alone will make Gaia one of the largest sources of spectra (Sloan will only have 1 million when it is finished). The instrument has an arcsecond positional accuracy.

2.4 Scientific Benefits

A search for Gaia on astro-ph yields 133 papers (May 2005) covering a broad range of scientific topics. Gilmore et al. [14] summarise the prime scientific possibilities with a thousand million object photometric and kinematic survey:

- The history of our galaxy: test hierarchic formation theories, inner bulge/bar dynamics, disk/halo interactions, dynamical evolution, what is the warp, star cluster disruption, weigh spiral structure, star formation history, chemical evolution, link to high redshifts.
- Stellar evolution: detect rapid evolutionary phases, quantify pre-main sequence evolution, complete census of local neighbourhood.
- Stellar formation: dynamics of star forming regions, luminosity function for pre-main sequence stars.
- Brown dwarfs: census of brown dwarfs in binaries.

- Planetary systems: complete census of (Jupiter size) planets around $3x10^5$ stars.
- The Local Group: rotational parallaxes for Local Group galaxies, kinematic separation of stellar populations, galaxy orbits to give cosmological history.
- Beyond the Local Group: parallax calibration of distance scale, zero proper motion QSO survey, photometry of 10^8 galaxies.
- The nature of matter: galactic rotation curve, disk mass profile from velocity dispersions, internal dynamics of Local Group dwarfs.
- Fundamental physics: determine the space-curvature parameter ? ? to 10^{-6}.
- Reference frames: define the local metric.
- Serendipity: the first all-sky, multi-colour, multi-epoch phase-space map.

A more comprehensive scientific case (100 pages) for Gaia is laid out in the Concept and Technology Study report (the red book) [13]. The Red Book is the document which culminated Phase A of GAIA before it was chosen as ESA's fifth cornerstone mission.

2.5 Gaia Processing

The author has long been interested in the global processing solution [22] for Gaia and indeed contributed a section [13] to the red book on the topic. Gaia processing differs from other astronomy missions because of the instrument design. To achieve the microarcsecond astrometry required to provide the scientific bounty outlined in [13] a rather complicated statistical processing must be carried out on the data.

Consider that the field of view for Gaia is about 0.4 degrees [26] consisting of pixels 44.2x132 milliarcseconds in size. With centroiding capability of 100^{th} of a pixel this gives a rough accuracy of 1 milliarcsecond for the instrument. Add the positional uncertainty of the satellite itself to this and it becomes clear there are issues to be addressed. Gaia offers the community accuracies of about 24 microarcseconds for objects G=15 and brighter. Each object will be observed about 100 times [28] on average. By constructing a model for the

stars position based on the multiple positions observed over the mission and the satellite pointing (and other factors) a more accurate estimation of the observational parameters may be computed. In a similar way the positions of the multiple objects observed by the pair of astrometric telescopes over time may be used to build a more accurate representation of the satellites attitude[2]. This is a feedback system, where improving one measurement improves the other – the intention is to iterate over these solutions until convergence is reached. This is of course an iterative and distributable solution for the otherwise intractable problem of solving millions of equations for the astrometric parameters of millions of stars.

The iteration of the algorithms requires access to the data in the spatial and time domains. Some of the algorithms require access to all observations of a given object e.g. astrometry and photometry calculations, other algorithms require all the data in time series e.g. to reconstruct the satellite pointing accurately or to calculate chromaticity [29] or other calibration values over time. This process is referred to as the astrometric Global Iterative Solution (GIS).

After, or simultaneously with, the calculation of the astrometry the other parameters must also be calculated. Expert groups around Europe will deal with photometry, bright stars, variable stars and spectrometry to name but a few. Many of these tasks in turn rely on output from the astrometric GIS and each other. The precise organisation of this is currently unknown but it is clear a large integration task will be required to put the catalogue together.

2.6 Gaia Data Flow

The exact data flow for Gaia is not known and will probably not be clear until end 2006 or 2007, but one may try to surmise what this might look like. Figure 5 depicts a possible dataflow cored around a Gaia Main Database as foreseen by the author. A version of this data flow was presented to the Data Analysis Coordination Committee (DACC) at their first meeting early in 2005. The author is heavily involved in the complete definition of the Data Flow and is a member of the DACC.

[2] The satellites attitude is its current position in space – one may think of this as the pointing.

Here Gaia data is picked up at Cebreros or equivalent ground station and transmitted to ESOC (European Space Operations Centre). ESOC routinely archives all telemetry and generally is the distribution centre for it. One could consider getting data directly from the ground station but then if the station is switched it would also be necessary to switch the receiving system – ESOC is already well equipped for switching ground stations. In Figure 5 a pass denotes one period of visibility of the satellite which may vary from 8 to 11 hours approximately. This represents about 24 hours of data.

EGSE (Electrical Ground Support Equipment) or other simulation data may be injected in the system at various points. Currently the GDASSII (Gaia Data Analysis Software Study) tests are run with simulated data.

Either at ESOC or another location "First Look" analysis will be carried out. It is not clear if the real time data from Gaia during the pass will be transmitted independently to the ground and swiftly sent to First Look in an alternative stream to the data stored on board. If this were to be the case some near real time detections of interesting events could be performed. In any case First Look will produce the first milliarcsecond positions of observations from the Gaia Pass. After First Look some calibration is still needed for the photometric and spectroscopic readings. These are further refined by iterative processing and application of the global derived parameters such as chromaticity. After calibration the data for the entire pass may be stored in a Pass Database (probably in some form of file system)

Next the ingestor would ingest this data into the main database. This would include organising, checking and indexing the data, incorporating the current pass data into the entire mission data set.

Over the lifetime of the processing there are a number of iterations for the GIS and other tasks each updating the main database. It may be simpler to consider just two versions at any time. The current version from which one takes values and the next version to which one writes new values. This may be seen as a globally versioned database where tasks read V_n and update V_{n+1}. Since tasks will probably be distributed at multiple locations it may be better to consider an Integrator which takes all the updates and prepares the next version of the database. The time scale for such versions would probably be on the order of six months.

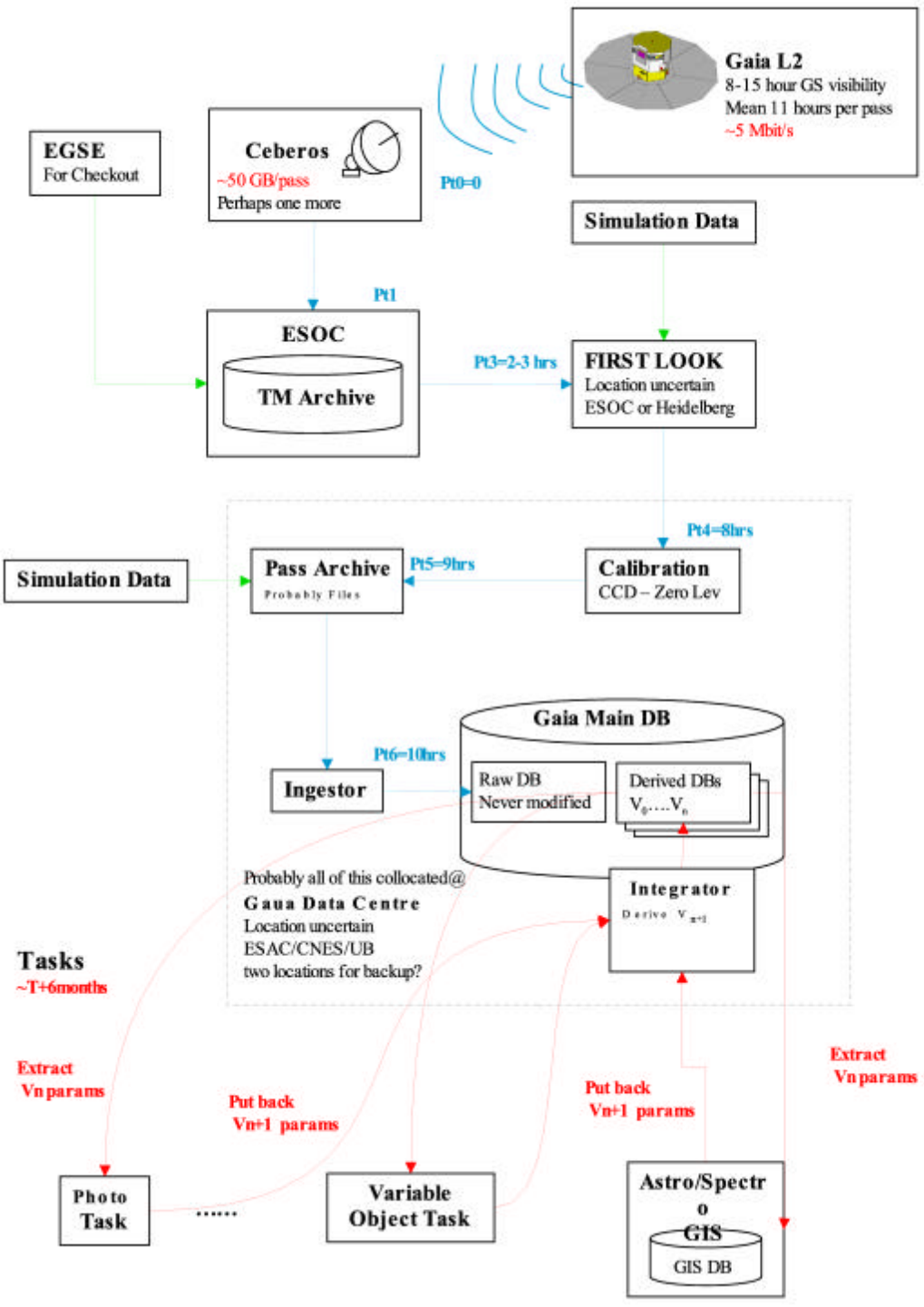

Figure 5. Possible data flow for Gaia.

The Main Database, Integrator, Calibration and Pass archives would probably be co-located in a Gaia Data Processing centre while the tasks could potentially be in any location which could manage a copy of the database.

3 Scientific Processing Systems

To support this study a series of interviews with managers in many large projects and institutes have been carried out by the author. The interview was based on a set of questions that were sent in advance to the interviewees. There were two separate questionnaires for missions and archives. The actual interviews were recorded and later a transcript was produced (these are not word for word transcripts), that was sent back to the interviewee to verify it was an accurate representation of the facts. In this section each of the projects or facilities are presented based on the content of the interviews and referenced supporting documents. Unfortunately not all these documents are in the public domain. Individuals came with various degrees of preparedness to the interview or simply did not have access to all of the requested information, which leads to a slight unevenness in the interviews. Some are obviously more biased toward technology others toward management.

In the sections below an attempt has been made to even this out and present similar information about the projects. Each of the interviews was very interesting and the transcripts contain many useful points. The transcripts are presented as appendices to this document

Five interviews were conducted with astronomy missions and two with astronomy archives. Some further non astronomy missions and institutes are mentioned but time and circumstance did not allow for interviews with relevant people at these locations to date.

3.1 Astronomical Systems

3.1.1 Sloan Digital Sky Survey

An interview on the SDSS [24][1] was carried out with Bill Boroski (Project manager) at Fermilab in Batavia Illinois on April 19th 2005 see Appendix 2. This was a very informative and open discussion with many interesting points raised in terms of management.

The SDSS is of interest to Gaia for its similarity in resolution and data volumes.

The Sloan Digital Sky Survey is a project to survey an 8000 square degree area on the Northern sky over a five year period (now to be extended for a further three years). A

dedicated 2.5m telescope is specially designed to take wide field (3 degrees in diameter) images using a 5x6 mosaic of 2048x2048 CCDs, in five wavelength bands, operating in drift scan mode. Additionally the telescope may be hand plugged with fibre to make spectroscopic measurements. The total raw data will exceed 40 Terabytes. A processed subset, of about 2 Terabytes in size, will consist of 1 million spectra, positions and image parameters for over 100 million objects, plus a mini-image centred on each object in every colour. The data is available to the public on one main website[3] with mirrors all over the world.

The processing and data management for SDSS is very challenging [24] a great deal of effort went in to its automation.

3.1.1.1 Mission Costs

The original five year survey was predicted to cost in the region of $25 million, the current cost to completion is estimated to be $85 million. The overspend was across the board for both hardware and software. Current running costs are $5.5 million per year to be reduced to $5 million per year for the additional 3 years.

3.1.1.2 Management

The SDSS team responsible for building the core data systems consists of ~180 people spread over seven main institutes with a further seven providing occasional input as well as funding.

SDSS is managed by a committee of four people: the director, the project scientist, the project manager and the project spokesperson (see Figure 1). An additional five level one managers form the entire management group. An attempt is made to collegially get everything done but if something is not going correctly management have learned they need to step in quickly and deal with it. At least two of the SDSS managers are formally trained in management.

[3] http://skyserver.sdss.org

It appears that early on roles were not well defined for individuals and it was difficult to hold people responsible for particular tasks. It is considered that many people involved were more familiar with working on small projects within their own sphere of influence and were ill prepared for an "industrial strength scientific project" like SDSS. This meant there was some resistance to the implementation of formal procedures in the project. Boroski points out that many of the team members are in Universities because they do not want to be burdened with formal systems and as such it is difficult for them to be in a large project which is necessarily quite formal.

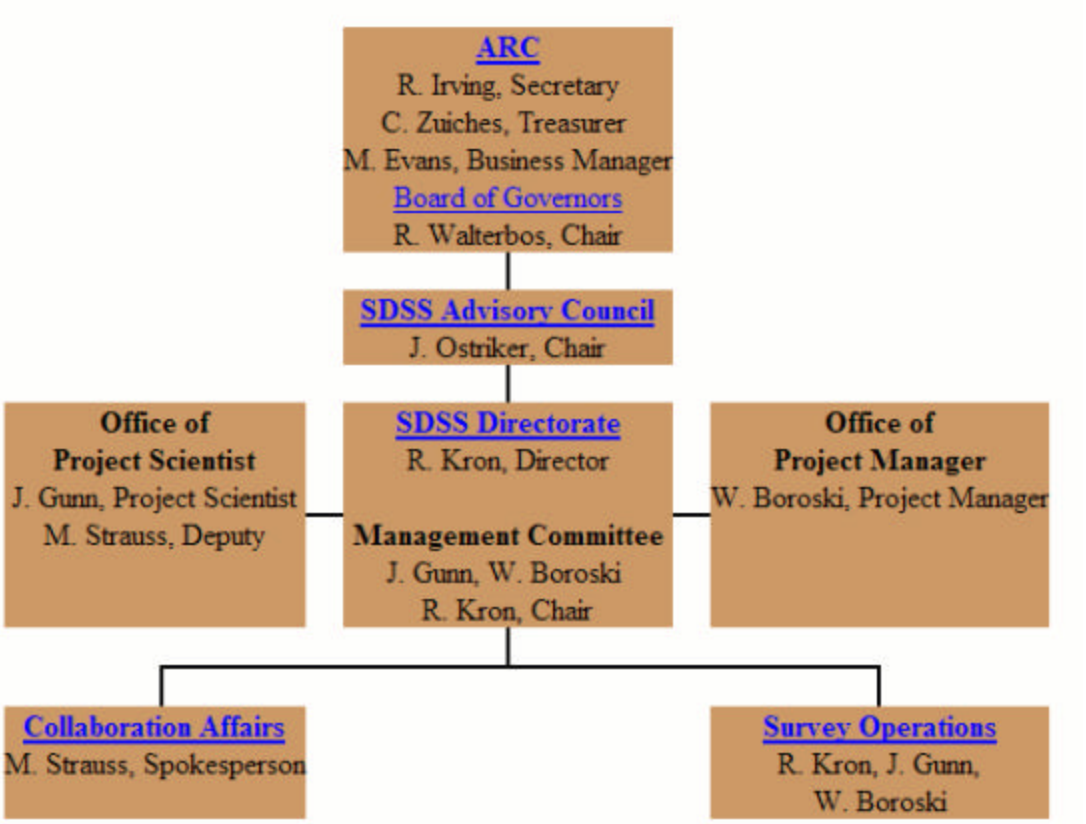

Figure 6. SDSS Organisation from the SDSS management page[4].

[4] http://www.sdss.org/directorate/index.html

3.1.1.3 Software

There were no formal software engineering standards adopted or mandated for SDSS. Some parts of the system development were carried out in a formal manner with requirements being written followed by implementation of the requirements. Other parts were built using a more holistic approach. There was also no particular design methodology for the code and many languages were used but most of the code is in C. Boroski notes there is a tendency of many developers to re-write rather than maintain existing code. There are no clear numbers but the estimate is that software cost a lot more than expected for SDSS. In retrospect Boroski says requirements should be hammered out with scientists early on and adhered to. Rules and procedures should also be put in place at an early stage.

CVS [6] is in use broadly across the system, all code is checked in to a central CVS system hosted at Fermilab. This has been seen as a very good idea even if it was initially difficult to get some people to use it. No release management software is used in SDSS, however a strong system is in place whereby developers tag code they consider ready for release, another party then checks this out and tests it before it becomes a formal release.

Gnats[5] is used for both Change Requests (CRs) and Problem Reports (PRs). For mountain[6] software there is a review panel to categorise and prioritise these, elsewhere this is not as well tracked as perhaps it should be.

Many software packages are used on SDSS ranging from DBMSs like Objectivity, SQLServer and MySql to Image processing packages such as ImageMagick. The feeling is these have always saved money.

3.1.1.4 Hardware

SDSS is definitely a distributed heterogeneous hardware environment. For processing purposes SDSS has over 40Terabytes of spinning disk. Storage Area Networks (SANs) were

[5] http://www.gnu.org/software/gnats/

[6] In SDSS software running directly at the observatory is sometimes termed “mountain software”.

considered too expensive so this storage is all connected directly to machines. Fibre Channel was found to be expensive and not particularly scalable – there is a move toward SATA[7].

A tape archive is used by SDSS. Tapes are sent from the mountain to Fermilab where they are loaded in a tape robot (ENSTORE[8]). This is how the data is accessed. A second set of tapes is sent after the first are received – these are put in cold storage. The tapes have been very reliable. The entire system is also backed up to ENSTORE. This is a Fermilab facility not something purchased by SDSS

3.1.2 Guide Star Catalogue

An interview was conducted with Brian McLean (Scientist) of the Guide Star Catalog (GSC) [18] group at The Space Telescope Science Institute (STScI) on April 13th 2005. He is currently in charge of the project. See Appendix 1.

GSC is for all intents a global survey based on old plates. It includes astrometry and proper motions. Its depth is similar to Gaia (18-20 magnitude) hence it has a similar size catalogue and range of objects.

The Catalogs and Surveys Branch of the Space Telescope Science Institute has been digitising the photographic Sky survey plates from the Palomar and UK Schmidt telescopes to produce the GSC and the Digitized Sky Survey (DSS). These catalogues support ground and space-based telescope operations and provide a valuable scientific resource to the astronomical community.

The Guide Star Catalog 2 (GSC2) is an all-sky catalogue based on 1" resolution scans of the photographic Sky Survey plates, at two epochs and three bandpasses, from the Palomar and UK Schmidt telescopes. The all-sky, magnitude-limited *Telescope Operations* version, GSC2.2, contains positions, classifications, and magnitudes for almost 1 billion objects, and is now available to the community via the WWW.

[7] http://www.serialata.org/

[8] http://www-isd.fnal.gov/enstore/

The management of the 8 Terabytes of scanned data and resultant databases was a complex task for GSC, the actual processing of the images was also a moderately complex task. Given the time scales it may indeed be quite comparable to Gaia.

3.1.2.1 Mission Costs

The original GSC/DSS project was part of the Hubble Space Telescope (HST) operations. A rough estimate of the original resources required for GSC/DSS were about $1 million for hardware and 100FTEs in manpower. It is possible that there was an overrun but it is not clear that this could be isolated from the main HST budget. The GSC2 project grew from GSC1 and external funding was sought. GSC2 cost to date is about $2million mostly in manpower (110FTES 60-70%). There were no overruns since the project was cut to fit the budget.

3.1.2.2 Team and Management

The GSC team fluctuated around the twenty person mark. In recent years the project has been winding down so now the team consists of approximately seven people. The team was distributed over three main sites: Space Telescope Science Institute (STScI) in Baltimore, European Southern Observatory (ESO) in Garching and the Observatorio Astronomico Torino (OATO) in Italy. Most of the team were based at STScI. The project was initiated and lead by Barry Laskar, he led the project by example and persuasion and not using any formal management style. The entire project and team were goal oriented toward the larger goal of completing the GSC catalogue. There was no formal mechanism for tracking progress for sub tasks, nor was there a very formal break down of them. There was little formal training for management in the project. Task estimates were generally well below the actual time needed to complete them, this was the case even late in the project. The main problem for management was the perceived lack of responsibility by the groups involved to adhere to the deadlines which had been agreed. There was a perceived lack of accountability on this front but it is unclear how this could be solved.

3.1.2.3 Software

There were no software engineering standards adopted or mandated for GSC. When Objectivity was selected for GSC2 there was a move to OO design and use of Rational Rose. This was not pervasive, Fortran processing software was still running on VMS in 2005. It was only with the movement of some code development to the Windows system that a source code control system was introduced. The chosen product was Visual Source Safe (VSS). However not all code was put in the system, only the code developed on Windows was controlled in VSS. The major off the shelf software for the project was the Objectivity database, for McLean it is unclear that this saved the project time or money. It is perceived that it changed the way people worked. Others on the team consider Objectivity to have saved time and effort on GSC2 and that perhaps it would not have been possible with out it.

There was no dedicated language for GSC. It started in FORTRAN moved into C++, had a little Java and is finally moving into C#. Code in all languages still exists. IDL has been in use throughout the project.

3.1.2.4 Hardware

As with the software, the hardware for GSC evolved over time. The original scanning machines were the major hardware cost. The image processing for the scans took place on a VMS cluster which is still operational today (2005). The actual catalogue production and further processing was migrated eventually to a large Windows server. GSC currently have about 25 Terabytes of spinning disk and are migrating away from their tape archive for reliability reasons. GSC have not chosen a SAN solution for cost reasons.

3.1.3 Large Synoptic Survey Telescope

An interview about LSST was conducted April 26th 2005 in Tuscon Arizona with Jeffrey Kantor LSST Data management project manager see Appendix 3. Kantor is an experienced manager and has interesting things to say about science projects.

The Large Synoptic Survey Telescope (LSST) is a proposed ground-based 8.4-meter, 10 square-degree-field telescope that will provide digital imaging of faint astronomical objects

across the entire sky, night after night. In a relentless campaign of 10-second exposures, LSST will cover the available sky every three nights, opening a movie-like window on objects that change or move on rapid timescales: exploding supernovae, potentially hazardous near-Earth asteroids, and distant Kuiper Belt Objects. The superb images from the LSST will also be used to trace the apparent distortions in the shapes of remote galaxies produced by lumps of Dark Matter, providing multiple tests of the mysterious Dark Energy [9].

LSST will pull down 25 Petabytes in its lifetime, 13 Terabytes per night of operations i.e. far more than Gaia. Each image will be around 3.5 GigaPixels. The aim is for 1% photometry[10] and arcsecond astrometry.

The processing for LSST is extremely challenging, an image the size of a DVD will be pulled off the telescope every 15 seconds and needs to be processed within a minute to satisfy the alerting requirements of the mission. The sheer volume of data is overwhelming compared to any other astronomy mission at the moment and its management will be extremely difficult.

This is definitely a sister mission that Gaia should keep in touch with. First light for LSST is due 2013.

3.1.3.1 Mission Costs

Currently in R&D phase (similar to ESA Phase A [11]) which is budgeted to between $24-28 million. The construction phase is budgeted at $270million with operation over 10 years running at $20 million per year. This puts LSST almost on a par with Gaia which runs to 500 million euro.

[9] From http://www.lsst.org/lsst_home.shtml

[10] Photons counted by each detector are accurate to within 1% of the "real" value.

3.1.3.2 Management

For data management two people are already in place. The project manager and project scientist, the expected peak team size is 15-16. This will be organised as in Figure 7. A charter document [17] exists with lists of responsibilities clearly defined for each role. A formal method for estimating cost was used for sizing the team based on COCOMO, FPA and SLIM/QSM. There are a large number of volunteers involved. There are currently 12 members of the consortium (all in USA) organised into three research teams. There are six working groups involved in defining requirements.

A democratic project management style is adopted. Kantor stresses though that once a decision is made he will not overturn it easily. There is a great deal of training and experience in the management team of LSST.

Kantor already finds working with an academic element somewhat difficult. As he puts it "Volunteers are definitely a two edged sword. You get some value but you really have to work at it!"

3.1.3.3 Software

LSST will use the ICONIX[11] process which is a trimmed down version of the Rational unified process. This seems to be a practical approach to get near extreme programming [3] while keeping some methodology in place. The methodology was chosen by the pooled experiences of the team. Some reporting standards are being adopted and it is expected they will be mandated by at least one of the funding agencies. Kantor feels the standards may be useful as a reporting tool if implemented properly and become more necessary the bigger the project. He also feels they may be a burden when a lot of documentation is needed.

LSST are using CVS [6] as a code repository but are looking into Subversion [25] they want a release management system but not something too involved. They will eventually put in

[11] http://www.iconixsw.com/

place a tracking system, but for now have an issue tracking system which is part of MS project.

LSST will develop in C++ and Python mainly. They wish to use off the shelf software where possible. Kantor sees a three tier system where the bottom hardware/infrastructure layer is 99% off the shelf. The middleware layer will probably also be mostly off the shelf, Condor, MPI etc. The Application layer will be mainly custom.

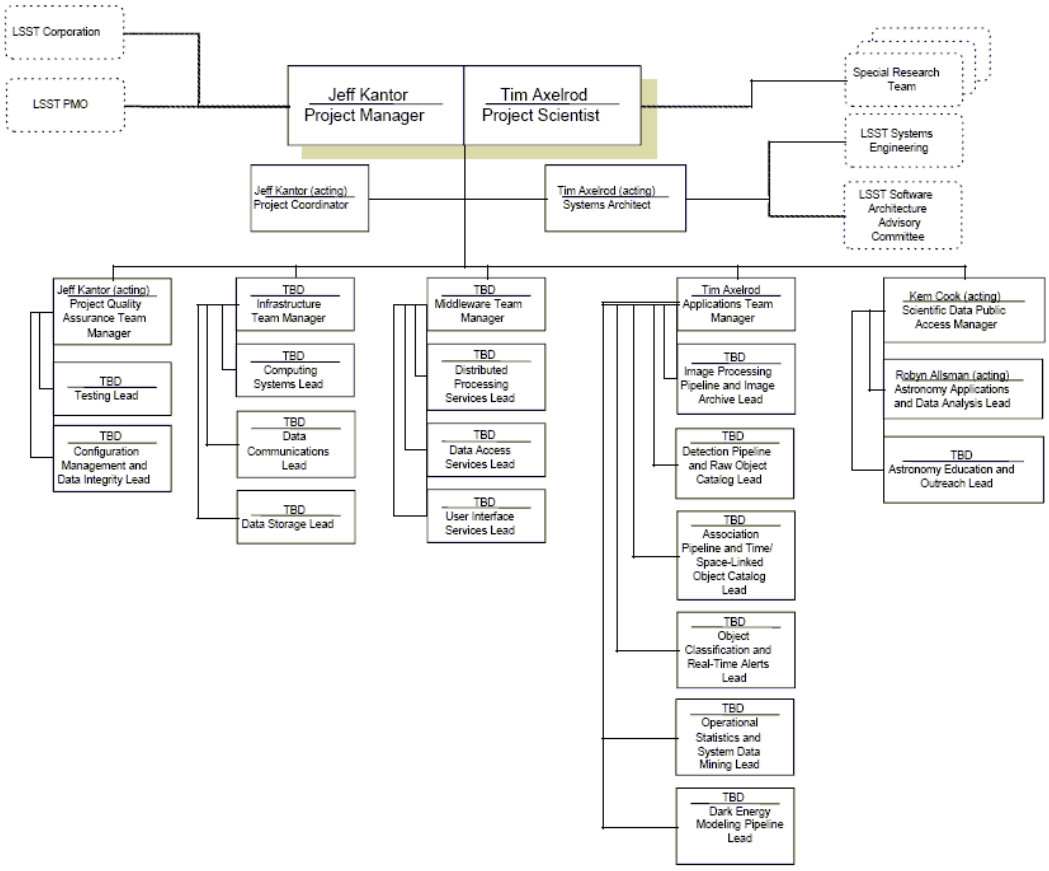

Figure 7. LSST Data Management Organisation

3.1.3.4 Hardware

LSST envision three different types of computing centres. The mountain base will have the data acquisition and pipeline servers as well as storage for two to five days of data. They

assume availability of at least a 2.4 gigabit link to the archive centre. This will allow for transmission of only the raw data– it will be cheaper to buy more processors than to send the processed data over the link. Yes you did not read incorrectly LSST intend to process on the mountain then **not** transmit the result but only the original raw data, which is smaller and must be sent in any case. The archive centre, then, will have a full pipeline system as well as data access servers for alerting and data distribution. There will also be data access centres that just give access to the data.

The estimate for processing power between the mountain and the Archive centre is ~75 Teraflops. With an estimated total of 30 Petabytes of data LSST feel they may need two or three times that amount of spinning disk. They are considering tape but feel it is a close call at the moment – all parts of the survey are important so it is not clear what would go on tape.

3.1.4 Planck Survey

Planck[12] is interesting to Gaia from the perspective of being a global survey with complex processing. Also it is a project Gaia may learn form in the way it has been managed. An interview was conducted with Jan Tauber (Project Scientist) on June 13th 2005 at ESTEC in the Netherlands.

Planck is an accepted medium sized mission and is due for launch in 2007. Planck will provide a major source of information relevant to several cosmological and astrophysical issues, such as testing theories of the early universe and the origin of cosmic structure.

The angular resolution of Planck will be 10 arcminute. Two instruments will be on board to give frequency coverage of 30-850 Ghz. The temperature sensitivity of Planck will be DT/T ~ $2x10^{-6}$ in the channels where the Cosmic Microwave Background (CMB) is the dominant signal and as close to this value as technically possible in all other channels.

Planck will produce nine complete maps of the sky in different frequency ranges. The raw data produced by the instrument will amount to around 0.5 Terabytes. In the words of Tauber

[12] http://astro.estec.esa.nl/Planck

a highlight would be to measure the polarized component of the fluctuations in the CMB. Planck is scheduled for launch in 2007.

The processing for Planck is conceptually not that difficult although computationally it is very challenging. Approximation methods have now been selected which essentially allow a best approximation of the map making and power spectrum extraction which is resource limited. Other systematic effects which were considered trivial earlier are now seen to be major problems.

3.1.4.1 Mission Costs

Planck is a medium sized mission and had an initial envelope of Euro 350 million. Various institutes will contribute a large amount also totalling around Euro 250 million. Not long after the selection of Planck (then COBRA/SAMBA), around 2000, ESA began studying the merger of Planck and Herschel (then FIRST). This was to reduce costs – the ESA envelope for the combined mission is now around 1.1 billion Euro.

The initial estimate for software development of 300 – 350 man years of effort was scaled from previous missions such as Hipparcos and ISO. It is essentially cost limited. The hardware (for processing) budget is not clear at the institutes but is in the order of a few million Euros.

3.1.4.2 Management

The Plank Data Processing Centres (DPCs) are essentially academic and the exact manpower for Planck is not yet clear. Many people are working a very small fraction of their time. The plan is to have around fifty people at each DPC as launch approaches, of these at least ten would be core full time Planck staff. The total scientific staff in each consortium is around 300 however spread over 50 institutes.

The style of management in the two DPCs is quite different. The Italian consortium tends to have a ridged hierarchical structure with a fixed core staff and loose affiliations with the scientific groups. The French consortium tends to have a looser management style relying

more on good will and judgement of the individuals rather than strong direction. The large British component in the French consortium also makes management difficult.

The managers in the consortia are at least trained informally and have a more management rather than science background.

Dr. Tauber points out that motivation seems to be greater in the less structured environment. He also feels that having a single consortium rather than two fairly independent ones would possibly be better. Managing the consortia has been very difficult and ESA opting to not have a major say in matters has not worked out for the best – it would be preferable for ESA to control or at least have a major involvement in the processing.

3.1.4.3 Software

Both consortia agreed to adopt PSS-05 Lite [12] when ESA mandated the use of some standard for software development. This has not been well adhered to however. Recently the DPCs have been refocused on launch critical software and this will be more formally checked.

There is no dominant methodology in use for software design although there is a structured approach to prototyping (bread boarding) and releasing software. Even that is not necessarily well adhered to.

CVS is in use fairly widely and seems to be doing well. There is a problem tracking system in particular for IDIS (Integrated Data and Information System) [5] but it is not used widely outside of that part of the project.

The DPCs have looked at Versant but ran into many problems with it, this could of course be due to a lack of expertise. The Italians may use Oracle for their data needs but currently are using files. The French are looking at Berkley DB[13], a free system which also has a java implementation. In the early days an interface layer was posited for access to the data at the

[13] http://www.sleepycat.com/products/db.shtml

two DPCs, this has now become smaller but allows the Process Coordinator (which controls pipeline processing) to interact with data from both DPCs.

The language is not fixed for Planck there is a lot of C/C++. Java has not proved popular.

3.1.4.4 Hardware

The Italians have a Beowulf system with 16 CPUs and they want to go to 30. The French have a large parallel machine and England has the Cosmos[14] supercomputer. There is a possibility to use the Grid for scientific exploitation but they wish to avoid this for core processing.

The initial hardware requirements were so huge for Planck that resource limited best effort analysis seems the only approach to the processing.

3.1.5 Integral

Lars Hansson, Integral Science Operations manager was interviewed at the European Space Astronomy Centre (ESAC) in Spain on May 27th 2005. The ESA scientific mission INTEGRAL (The International Gamma-Ray Astrophysics Laboratory) is dedicated to the fine spectroscopy (E/deltaE = 500) and fine imaging (angular resolution: 12 arcminute FWHM) of celestial gamma-ray sources in the energy range 15 keV to 10 MeV with concurrent source monitoring in the X-ray (3-35 keV) and optical (V-band, 550 nm) energy ranges. The processing of the gamma ray detections is fairly difficult.

The ground segment consists of two parts: the uplink and the downlink. The downlink part, data processing and distribution to the community is done by a PI consortium at the Integral Science Data Centre (ISDC) under the Observatory of Geneva. The uplink part, Integral Science Operations Centre (ISOC), is done by ESA with all software developed in house under the direction of Hansson. Initially ISDC also interacted with ISOC but now in

[14] http://www.damtp.cam.ac.uk/user/gr/cosmos/science/science.html

operations there is triumvirate of ISDC, ISOC and ESOC (European Space Operations Centre).

3.1.5.1 Mission Costs

Integral is a mid-size mission with an ESA budget of 400-500 million Euro. The ISDC and science community have some independent funding but ESA also contributed a good deal to the community effort (at least 20 man years). To date Integral has remained within budget overall, investing a little more in the instruments than originally planned. About 40 man years was budgeted and used for ISOC development so it was within budget although there is a feeling that without the launch delay they may have gone over. ISDC has used about 200 man years to date – it is unclear if they had an actual estimate at the outset. ISOC has currently spent about 300K Euro on computer hardware but it is insufficient. They are hosting a copy of the archive which has turned out to be bigger than planned with many more products than foreseen. They have been saved, to a degree, by the falling price of disk.

3.1.5.2 Management

ISOC has had 4-6 people working since the outset and ISDC has had 25-30 people. ISDC has collaborators in around ten institutes including non-EU institutes. ISOC has been fairly tightly managed in ESA style from the outset with a development manager reporting to the ISOC manager and the team reporting to the development manager. Later the ISOC manager took on the development manager role also. ISOC followed ESA standards and so have a good deal of documentation including the Software Project Management Plan which include the organigram shown in Figure 8.

ISDC has been a little more fluid in their approach to development with a more academic feel to it. It has been ensured that some good procedures and configuration management have been implemented there.

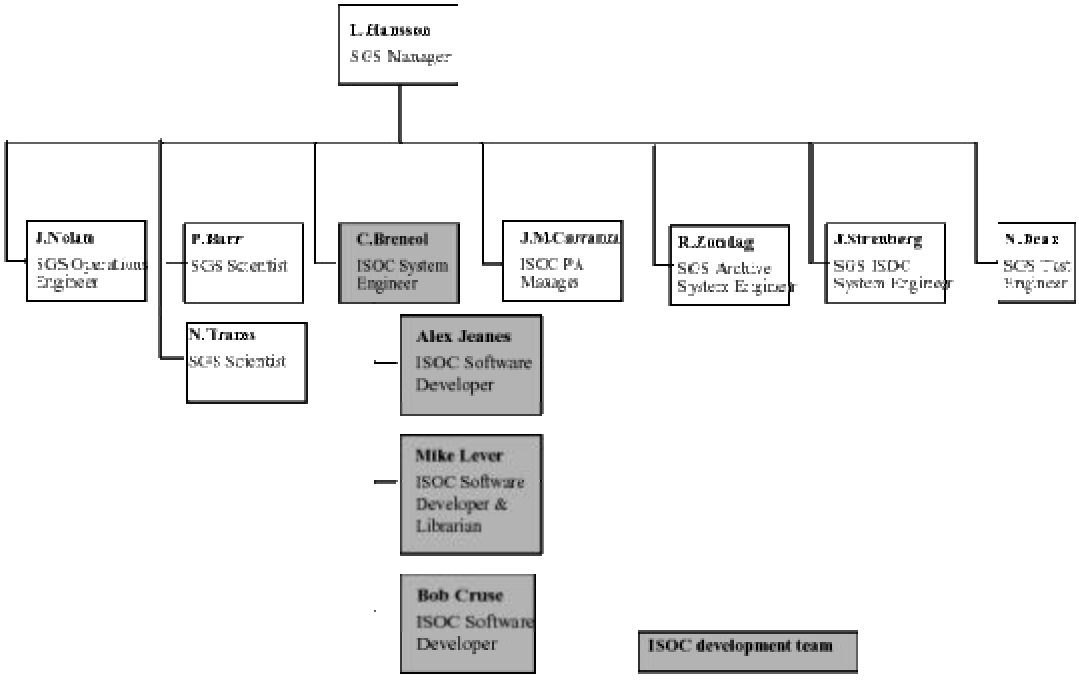

Figure 8. ISOC Management (from Soft. Project Management Plan).

Figure 9 shows the operational organisation of the ISDC with eight distinct teams reporting to an operations manager. Operationally Integral needs to report near real time GRBs (Gamma Ray Bursts) and requires an operational level of reliability. This is working well now and performing nominally.

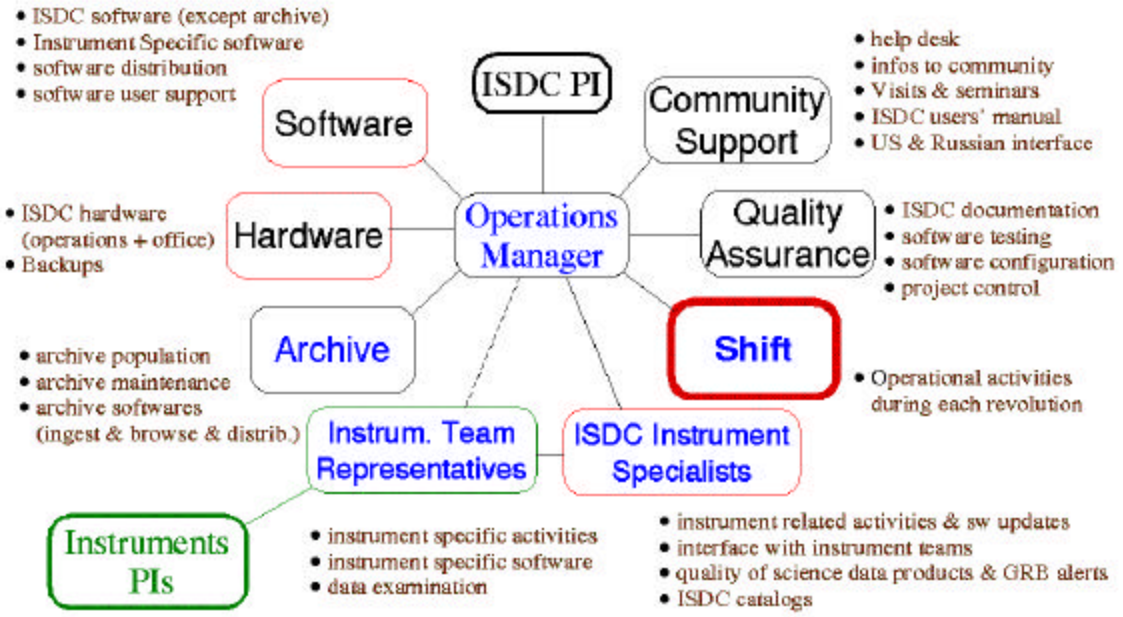

Figure 9. ISDC Organisation from [21].

3.1.5.3 Software

ISOC used the agency standard for software engineering PSS-05. The agency switched to ECSS during the mission but existing systems were not required to refit their documentation to the new standards. These were felt to be useful and strict configuration control is still maintained (although some of the scientists still grumble about this). Rational Rose and a unified like process were used in ISOC initially but this was relaxed somewhat after the ADD production as keeping the ADD in sync was considered too time consuming. The tools were selected based on team experience and outside consultation. This may also have been partially to do with individual ability. Both ISOC and ISDC have had a QA presence from the outset. This was felt to be a good benefit and led to no problems for ISOC in the launch readiness review.

CVS is in use and found to be beneficial – no special release system is used beyond tagging in CVS. A problem tracking system was taken over from ISO but there is a much better system in house at ESAC now.

ROOT[15] /C++ is the system of choice at ISDC and they have partnered with CERN on that. At ISOC Rose, Oracle and Java/JBuilder are used. Junit is used for unit testing in ISOC. Hansson points out that having a good team is all that really counts and he is very happy with the Oracle/Java choice.

3.1.5.4 Hardware

At least one Solaris machine is needed in ISOC to run the Flight Dynamics software that is only available on Solaris (ESOC have not released the code). Initially ISOC was developed on Suns but they have moved to Linux/PC for cost reasons. Now Sun are coming back with cheaper machines also. ISDC also developed originally on the Sun Platform but supported PC/Linux machines. ISOC currently has 6 Terabytes of spinning disk which was the total estimate at mission start. It now looks like they will need at least 12 Terabytes.

3.2 Astronomical Archives

3.2.1 Centre De Donnes Astronomiques de Strasbourg

An interview was conducted with Francoise Genova the Director of CDS Strasbourg who was at the time in Kyoto at the international conference centre on May 19th 2005, see Appendix 4. CDS was founded in 1972 by the Intitute de Astronomie et Geophysique to look after ground based facilities. CDS performs several roles including providing reference and value added services to the astronomical community. The CDS serves the Hipparcos data as well as numerous other larger optical and infrared surveys such as Guide Star Catalogs, the USNO-B1, and the 2MASS last release through its Vizier interface. Simbad is a world renowned service provided by CDS which resolves names and finds references for objects in astronomical journals. CDS plays a major role in the International Virtual Observatory Alliance (IVOA) leading the Unified Content Descriptor (UCD) definition and participating in practically all working groups. Aladin, a sophisticated visualization tool for astronomical images and catalogues, is provided by CDS and used in many VO applications around the

[15] http://root.cern.ch/

world. Francoise Genova was appointed deputy chair of the IVOA in Kyoto in May 2005, the deputy chair becomes the chair the following year.

3.2.1.1 Mission Costs

The main costs at CDS are for personnel and these are typically government appointed positions. Hence the overall costs are difficult to assess. CDS provides reference services, as such many of the staff are librarians, some are astronomers and there are a few software engineers. Hardware is a small fraction of the overall costs.

CDS has many projects on going and adopts a global strategy of adjusting effort to projects depending on their perceived pay off in long term usefulness to the astronomical community. There are few tightly budgeted projects making it difficult to assess overruns in project planning. Some projects have taken longer than expected.

CDS cut their projects to fit their allocated staff and monitor their progress closely.

3.2.1.2 Management

It is important to note that CDS was put in an observatory from the very beginning to ensure scientific expertise and to keep the focus on serving the astronomical community not on building technical tools. CDS started off with just a few people and had only one software engineer for the first fifteen years. It grew steadily from then on and now there are 25 to 30 people working on CDS projects spread over five locations but with the majority in Strasbourg,

There is no formal training for managers in CDS. The CDS mission is complex and involves following trends in at least astronomy and technology. The director must balance the needs of the individual projects against the overall impact of CDS on the community and this is done with a relatively fixed compliment of staff. Consensus is built in meetings involving people with a broad range of profiles.

Genova analyses failures in CDS but does not see that any major change would have facilitated a better global outcome – a different outcome perhaps but not necessarily globally

better. The global perspective of providing quality services to the community pervade and individual failures are perhaps less important in the light of overall success.

Although not always possible, Genova feels it is important to tailor management to fit the style of project being undertaken, the goals of the project and the constraints of other organisations which are involved. Goals and roles should be clearly defined and agreed for each organisation.

3.2.1.3 Software

CDS use no software engineering standards and no particular software engineering methodology. There is no formal bug/change request tracking system. Software products are chosen on a per project basis according to suitability. They have Object Oriented (OO) and Relational Database Management Systems (DBMS) as well as in house file based systems. The in house developed systems are being migrated to POSTGRES[16], which is a free system, to facilitate long term maintenance.

The main language at CDS was C/C++ but Java is now in use by many people.

3.2.1.4 Hardware

CDS runs a heterogeneous environment with workstations and PCs running Linux. There is roughly 5-6 Terabytes of disk spinning in CDS currently. They previously had a main frame and Simbad had to be moved from that platform to Unix. They are moving away from tape for backups opting rather for a complete copy of everything on disk in another building.

3.2.2 High Energy Astrophysics Science Archive Research Center

An interview was carried out with Tom McGlynn Chief Archive Scientist of the HEASARC for NASA on July 5th 2005 at Johns Hopkins University. NASA provides a set of archives in wavelength domains and provides the long term storage for the data in those domains. The HEASARC serves up online the results of many high energy missions including ROSAT,

[16] http://www.postgresql.org/

ASCA, BeppoSAX, Chandra, Compton GRO, HEAO 1, Einstein Observatory (HEAO 2), EUVE, EXOSAT, HETE-2, INTEGRAL, Rossi XTE, and XMM-Newton. In the coming years they will also provide the interface to the Swift Gamma-Ray Burst Explorer (2004), Astro-E2 (2005), and GLAST (2007).

3.2.2.1 Mission Costs

The HEASARC budget runs to about $4 million per year which is mainly for personnel, approximately half of the effort is in software development and half in scientific support. As little as $100k per year is spent on hardware.

3.2.2.2 Management

There is no particular management style in the HEASARC and managers may be trained but it is not mandated. The staff report directly to the head of the HEASARC (Nic White) there are a mix of contractors and staff with scientific and technical staff coming from different contract agencies to keep control completely in the hands of NASA. This is not necessarily seen by all to be a good idea. There are between fifteen and twenty people working at the facility.

Between existing and projected missions the HEASARC interacts with about 10 missions each of which is associated with some institution. Apart from data acquisition however there is little dependency from the HEASARC on the institutes.

McGlynn feels that the initial decisions on work processes are very important and that it is very difficult to change these once they are established.

3.2.2.3 Software

There are no formal standards for software at the HEASARC, the main package, FTools, does have a formal set of regression tests and release policies. There are five million lines of code in the HEASARC which grows organically as features are added and bugs are fixed. CVS is in use for most but not all of the code but no release management software is

employed. Bugzilla[17] is used for problem tracking as well as a home grown tool for the FTools, however the majority of bugs do not go through the system.

A host of COTS (Common Of The Shelf) or free software is used at HEARASC. Nagios[18] is used to measure uptime of the system, Sybase is used for the data storage which is slower than the old proprietary system since transactions are not required. McGlynn feels it saves money but perhaps not as much as one might think.

There is no particular development language: Perl Tcl, C, Fortran and Java are all used.

3.2.2.4 Hardware

The hardware varies at the HEASARC but mostly it is PCs running Red Hat Linux. For the archive the estimated power is about three or four Gflops. There is about twenty Terabytes of disk in SAN units, about ten of that is user space. tapes are used for backups.

3.3 Non astronomy science systems

There are a number of other science data/processing centres and projects of interest. Particle physics experiments in particular produce far more data than Gaia will need to deal with, although they throw much of this away. Dealing with huge volumes of data is of interest to Gaia.

Unfortunately time has not permitted interviews of further studies of the following systems which are of interest.

3.3.1 CESCA

In Barcelona CESCA support many project in different science fields. The current GDASS study code runs in this facility. They have a mix of high end machines and storage systems

[17] http://www.bugzilla.org/

[18] http://www.nagios.org/

but are beginning to look at cheaper Beowulf type architectures. Gaia has contact with this facility in any case through UB.

3.3.2 BSC

The Barcelona Supercomputer Centre (BSC) have recently built a new super computer 'Mare Nostrum' [33] from off the shelf components in a short space of time. This endeavour made news around the world and highlights some possibilities for Gaia. It would be interesting to compare this to buying commodity super computers such as offered, for instance, by the Orion company.

3.3.3 BaBar

Building on the original 2200-meter PEP storage ring and in cooperation with LBNL and LLNL, SLAC is constructing an extensive upgrade called the B Factory which will produce millions of B mesons. This upgrade includes modifications to the PEP storage ring and a new type of detector, called BaBar[19].

The BaBar science system was initially built around Objectivity and was one of the first projects to write several terabytes to Objectivity managed storage. From looking at their web pages they are now using ROOT in part of their system but still purchased Objectivity licenses this year. Reference is also made to this transition in [4].

3.3.4 Large Hadron Collider

The LHC is an accelerator which brings protons and ions into head-on collisions at higher energies than ever achieved before. This will allow scientists to penetrate still further into the structure of matter and recreate the conditions prevailing in the early universe, just after the "Big Bang"[20]. There is an interesting paper all about managing the petabyte of information from LHC [4]. They have used Objectivity successfully and are now looking at re-

[19] from http://www2.slac.stanford.edu/vvc/detectors/babar.html

[20] from http://lhc.web.cern.ch/lhc/general/gen_info.htm

implementing the system differently. Choosing a different approach brings different problems.

4 Application to Gaia

Following the broad outline of the related projects in Section 3, here a discussion of management, software and hardware for Gaia is presented.

4.1 Mission Costs

ESA costs for Gaia are on the order of 500 Million Euros. There is no estimate yet for the additional community commitment to the project but one may assume several million. LSST, Planck and Integral are of similar scale on cost basis. Taking into account the high cost of satellite missions e.g. the launch alone takes a large percentage of the cost, SDSS could also be considered similar. From a data volume perspective LSST will have far more data, although more traditional processing, than Gaia while SDSS is similar in size (considering it covers only part of the northern hemisphere). Hence based on cost this is a good group of missions to look at.

4.2 Management

In mid 2005 the Gaia Data Analysis Consortium Committee (DACC) was formed with the mandate to organise the community for the processing/reduction of the Gaia data, a daunting task. This will be where the rules are laid down for the future of the group. Almost every project manager interviewed expressed the opinion that whatever is set up in the beginning sticks and becomes very difficult to change so this needs to be done properly from the outset.

A relatively formal management approach should be adopted from the outset. Roles should be clearly defined and agreed for all of the major players– this is in fact practically mandated by the ECSS (European Committee for Space Standardisation) standards [8]. ECSS are the standards of choice for ESA. Starting out formally may allow some relaxation of norms later whereas the reverse is much more difficult to achieve.

The major problem of scientific projects is the lack of accountability of groups in the consortium. In the Gaia project interdependency between groups needs to be minimised while dependencies which will lead to catastrophic failure need to be highlighted to the group. The

groups involved in the processing need to be acutely aware that any deviation from the plan may cause the entire enterprise to fail. The productivity of the groups needs to be carefully measured, this does not necessarily mean how much they produce but rather that they produce what they agreed to produce within the time they agreed to produce it. Extreme programming [3] has an interesting 'points' technique for grading programmers based on their estimates and actual achievements. Some formal system such as this should be put in place project wide. The development of such a system is beyond the scope of this document however.

Accountability of groups for any large scientific collaboration is difficult in part because of funding. In projects within industry management have the enticement of bonuses and the fear of dismissal as 'hard' tools to motivate staff. The management have fiscal control over the project and the employees. In a scientific endeavour, such as Gaia, most of the work is performed on a collaborative basis with each group sourcing and controlling their own funding, hence the accountability does not lie with a central management team, rather it is more disbursed. The management team of the science project are left entirely with 'soft' management skills to achieve the project. This makes science projects far more difficult to manage than industrial projects and yet the 'managers' of science projects are seldom trained at all in management techniques. In what may be considered the bible of organisational management [16], even Handy has little to say on this type of project. In the projects above the level of training of managers varied. Some suggested more training for managers would be better. All Gaia managers should be sent on management training courses. What constitutes a manager for Gaia would need to be defined but initially it could certainly mean the Coordination Unit (CU) leaders and their seconds. An interesting idea may be to send all of the managers on a course together. This would give an opportunity for a diverse group to get to know each other and also form some ground rules within the group from the outset. It is common in industry to have team/leadership building courses but this is unheard of for scientific projects. As Genova points out it is important to mould management styles to the project, it will take time to put the Gaia management in place. A common course could precipitate a shared management style while forming a 'consensual domain' [20], a shared understanding of management terms, for Gaia.

4.2.1 Cost Estimation

Every science project, it seems, underestimates cost. This topic applies only to the processing since the instruments/satellite will be on a fixed priced contract with industry. Gaia is already relying on the falling price of hardware for computing and is indeed at risk of also underestimating cost. Most of the science missions are limited by cost – in some sense Gaia will also be cost limited, getting to do only the best possible job with the available resources.

LSST are using formal cost modelling methods to estimate manpower requirements for their system – such a model has not been developed for Gaia to date. It is not clear whether these models are actually better than the 'expert' estimate approach currently in use, Gaia should investigate some manpower modelling methods.

Cost estimates for scientific projects are arguably more difficult than for industrial projects. Usually the time scales are larger, the unknowns more copious and there is seldom a similar project to copy actual cost from. Again this begs for more formal training for some management entities in Gaia.

4.2.2 Organisation of the Consortium

How much rigidity and formality needs to be put in place? Tauber points out that a very rigidly controlled group may be less creative and motivated than a more loosely organised team. Most of the projects reviewed had very light management, but also felt they could do with some or slightly more management. The DACC are currently looking at a fairly hierarchical organisation with a committee sitting above seven Coordination Units (CU) of which each may have several development units beneath them. The exact definitions of these are still forming but could be seen as in Figure 10. Here not all of the Development Units (DU) are filled in – their number and composition is something that will fluctuate over time. Here the CU is really a management position which controls many DUs – not a composition of units, rather a control/management structure. CU managers report to the Data Analysis Consortium Executive (DACE), which ultimately answers to the Project Scientist with some input from the Gaia Science Team.

DUs do not need to be static; they could be created for the duration of a particular task and then disappear later. This is a very appealing idea. Dynamically composed teams for specific tasks would allow distribution of manpower through the consortium and its application where it is needed. Some tasks will require cross institution coordination, having a DU with members from two or more institutes may work very well in some cases. The tasks would be given to CU managers who would form DUs to perform the tasks – all manpower should be reconciled at the DACE level. CU managers should be collocated with the bulk of their DU manpower.

The Work Breakdown Structure (WBS) and Organisational Breakdown Structure (OBS) are not clearly differentiated in the current manifestation. The Work package breakdown should not be organised along the CU-DU line - it should be more architecture oriented. Broadly the CU is close to an architecture block but a more definitive architectural design is needed to ensure nothing is missed. If the two happen to align perfectly that would be fine. The Dataflow and various other pieces are a start for this design, but more is needed.

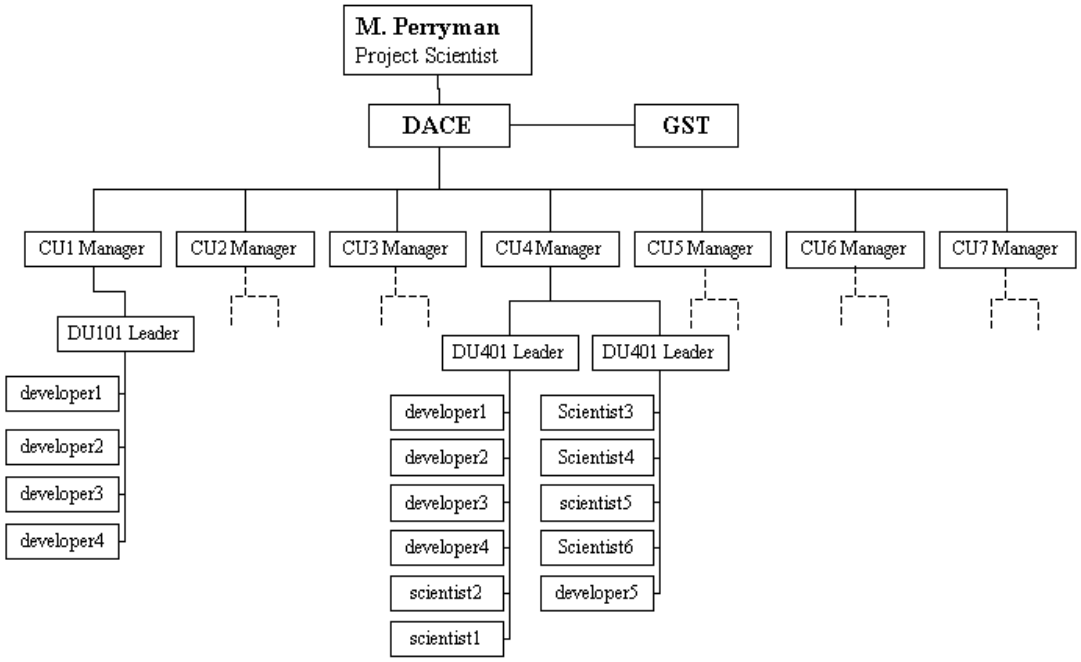

Figure 10. Possible Gaia Organisation

ESA are taking a large role early in the Gaia Processing which is good. The lack of a major ESA role in Planck was lamented by Tauber. ESA having a role in the processing give them 'hard' management tools over at least one group in the processing consortium. This group may be used to cut down interdependencies between the other groups by having them all interact with the ESA controlled group rather than each other.

4.2.3 Planning

The Gaia project as a whole is a huge undertaking running into the order of 500M Euro for ESA. The Scientific community will undoubtedly come up with a few more million for the data processing. Still the resources available are small compared to the task at hand and will take careful management. Currently the DACC (Data Analysis Consortium Committee) are in place to start this process. Nominally this would start by producing a functional breakdown, at least according to ECSS [7] (European Cooperation for Space Standardisation) others have a less rigid view [19] but there is some agreement to having some sort of model for the project. This is lacking currently in Gaia and needs to be addressed.

After a model is defined distinct phases should be defined. There are obvious points in the mission for this – pre-launch, operations, and final catalogue production. ECSS-M-30B [10] is probably not useful here as it is quite geared toward a complete satellite mission.

The current organisation into eight coordination units makes planning at least a remote possibility but it is still a difficult task. The management and standardisation of the work packages alone is complex – a more detailed discussion is in [23].

Several of the interviewees pointed out that whatever is put in place has a habit of sticking around – effort needs to be, and indeed is being, expended to get the planning as accurate as possible.

4.2.4 Standards and practises

Most of the interviewees felt they would have benefited or had benefited from utilisation of standards. Gaia is an ESA mission and will possibly be required to use the ECSS (European Cooperation for Space Standardisation) standards. This is already under investigation. These

are relatively new, having been adopted by ESA only in the last five years or so, some missions such as Integral received permission not to use the new standards. Hence there is less experience with them than might be desirable. But they are certainly desirable and Gaia should adopt and adapt ECSS standards for the processing system. The 'Lite' standard as outlined for PSS-05 in [12] should be investigated when tailoring ECSS. Furthermore some staff should be sent on ECSS training courses with a mandate to understand and tailor them for the Gaia data processing.

Many projects archive messages sent to mailing lists. Such a system keeps a track of all exchanges without the need for each participant to keep all of the emails. Many systems allow this to be browsed on web pages later making it readily accessible to new team members and allowing links to be made to individual messages for reference. The International Virtual Observatory Alliance[21] (IVOA) lists are a good example of this. Existing Gaia mailing lists should be moved to such an archived browseable system.

The author uses instant messaging and Skype quite often to assist others and to communicate with team members. This is a very nice, if informal, method for working. Some projects such as Astrogrid use Jabber in a more formal way to have organised meetings. Teleconferencing remains better for a remote meetings involving more than one party – but Instant Messaging could be formalised for Gaia for adhoc questions. It would be interesting to see the availability of key contacts in the Gaia team on an instant messaging system. A range of weekly teleconferences are held for SDSS and NVO to keep the project manager and team up to date on activities. Soon Gaia will also need regular teleconferences to keep on top of DACE activities.

Wiki[22] webs are becoming very popular and indeed Gaia already has one for the Data Processing. Wikies are good distributive collaboration tools and ideal for this type of project.

[21] www.ivoa.org

[22] www.wiki.org

4.3 Software

There is no clear trend in the survey projects concerning software. Most of the projects think they benefited from using commercial or public software where possible. Only GSC felt that the use of a DBMS changed their working habits and did not necessarily save them effort, and that is not a team wide opinion. The LHC paper [4] also points out that for complex data management each product brings its own problems, they made their system work with Objectivity and are making it work again with Root (it is not clear why they are re-writing a working system). Where possible Gaia should use off the shelf components and indeed this is the current practice.

Gaia should maintain its flexible outlook on DBMS software. The interface layer for data access needs to be conserved and the performance of different data systems tested. Such activities are in the current planning. It is interesting that Integral, and two major High Energy Physics Projects use ROOT, although they all have ties with Geneva. ROOT should be examined and contact should be taken up with CERN to learn from their experiences.

CVS for configuration control of the software seems to be clear across the board and one thing which all projects have in common apart from GSC. Even GSC use a configuration control system but it is Microsoft's Visual Source Safe, which is only available on windows platforms. Gaia already uses CVS in places and it should be brought in across the board as standard practice. Gaia should also have a formal and preferably electronic change request and problem reporting/tracking system, again such a system is under investigation.

Many of the projects bemoaned having insufficient requirements and specifications for software. As pointed out above Gaia should adopt ECSS standards to some degree. As well as defining requirements for all the Gaia deliverables it is important to focus on the essential tasks. This has already been raised at the DACC, there is a fine line between Gaia data processing to produce the catalogue and data analysis which may be of scientific interest but not strictly necessary for the catalogue production. It is good that this point has been raised early in the process.

There seems to be no clear indication on the programming language to use. The current approach of using Java for its portability seems good. It is interesting that LSST are moving straight into C++ but this may in large part be due to team experience. Current projects and experience show Java to be cost effective for development e.g. one is less likely to underestimate the development task using Java. Considering that most projects seem to underestimate tasks it should be advantageous to stick to a Java like language. Integral have been very happy with Java while it did not really take off on Plank. Gaia must consider the fact that there will be other languages used in the project as well as multiple DBMSs. It may be good to set initial fixed review times for the major Language and DBMS recommendations e.g. use Java and Oracle until 2007 at which point review available languages and DBMSs and confirm the decision. The current approach of portability and flexibility remains the best way forward for at least the next few years. The data flow presented above would allow different architectures/languages and DBMSs for individual parts of the system.

Finally as many of the tasks as possible should be automated, manual intervention is costly and error prone. There will be enough instances where investigation will be required when things do not go quite right, this will be time consuming enough without the need for day to day intervention. This should be extended down to the build system for the code, a product such as Cruise Control should be used to initiate automated builds. The Gaia satellite is designed to need little or no intervention for its five year mission, the ground system should strive to be self sufficient.

4.4 Hardware

It is early days for Gaia to look at hardware seriously. Most of the projects have built on cheap disk systems attached to machines rather than the more expensive Storage Area Networks (SAN) currently in use in some Gaia sites (e.g. Barcelona and ESAC). This is usually a cost decision and will be no different for Gaia in the long run. It does seem SANs and similar technologies are falling in price and a petabyte SAN in ten years may cost only about three million euros. At least for some of the Gaia sites this may be achievable – depending on how much is needed for processing power.

Apart from SDSS no one has a particularly good word to say about tape. The preferred method is to buy more cheap disks and keep all the data online. This can be expensive – disks cost money to spin and produce a lot of heat, however disks which spin down in a power save mode will surely come to market soon to alleviate the power and heat problems. It is clear that if tapes are to be used, they need to be high quality; they will therefore also be expensive.

Critical Gaia data should be kept in at least two locations. In the current scheme this will be the case as raw data will be shipped to at least one other site and all sites will have a copy of the Main Database (at least the current version). This does place certain hardware requirements on the sites.

5 Conclusions

Several interesting points were raised by conducting interviews with some of the major Astronomy projects of the moment. An attempt has been made to homogenise the information and apply it to Gaia.

Management of science projects is more difficult than industrial projects yet science managers tend to have less training in the management field. A more rigorous approach to management would help with accountability as well as cost estimation and ultimately achieving the overall project goals. Gaia needs to continue on its track of relatively formal management while formally training the managers in key positions for the data processing, namely the CU managers. Accountability needs to be ingrained in the CUs from an early stage, deadlines must be agreed with CUs and then they must be held sacrosanct. To reduce the possibility of failure due to under performance of any group involved in the processing, interdependencies need to be kept to a minimum. ESA taking a major coordination role here should help since it will be populated with ESA funded staff over whom ESA at least have fiscal controls. This ESA controlled group may form a hub for the processing dependencies and thus avoid, or at least reduce the complexity, of the interdependency network between the groups.

In addition to the WBS Gaia needs to define phases for the Gaia data processing project and carry out a more complete architectural design. Work packages resulting from the architectural design need to be assigned to CUs to ensure a full processing model is in place. The ECSS standards should be adopted for Gaia processing and at least some staff should be trained on the application of the standards.

Gaia is doing well on the software front; Java is a good decision for the foreseeable future giving portability and low cost software. However, the choice of language does need to remain on the table for later. The choice of Database Management System needs to remain flexible, as is currently the case. CVS is already in use and electronic issue tracking is in place. Gaia needs some browseable mail archives. Where possible Gaia needs to automate tasks from software builds up to data processing.

It is too early to make major considerations concerning hardware for Gaia. Developments in technology need to be monitored and costs considered later in the mission.

It seems Gaia is already doing many of the things which would be recommended by the managers/scientists interviewed. This is excellent news and shows that Gaia has a very experienced and dedicated team in place already. The road is long however and problems will surely arise. Gaia data processing, although a mammoth task and probably facing insufficient funding, needs to remain nimble in many areas – this is a challenge in common with many scientific projects of course, but each set of new projects are more ambitious than their predecessors amplifying such challenges.

Here a start has been presented which the author hopes will help in the formative years of the Gaia Data Processing project.

6 References

[1] Abazajian K. et al. (SDSS collaboration), "The Second Data Release of the Sloan Digital Sky Survey" Astron.J. Accepted. http://xxx.lanl.gov/abs/astro-ph/0403325

[2] Arenou F. & Babusiaux C, "Centroiding accuracy of bright stars", AAEB-FACB-02, V1.1 internal document (2002).

[3] Beck K., "Extreme Programming Explained" Addison Wesley 1999.

[4] Becla J., Wang D. L., "Lessons Learned From Managing a Petabyte" CIDR 2005.

[5] Bennett K., Passian F., Sygnet J.F., et al., "Sharing data information and software for the ESA Planck mission: the IDIS protoype" Proc. SPIE 4011, 1-10, 2000

[6] Cyclic Software, "CVS version control", https://www.cvshome.org/

[7] ESA Publications Division, "Project Breakdown Structures", ECSS-M-10B, 13th June 2003

[8] ESA Publications Division, "Project Organisation", ECSS-M-20B, 13th June 2003

[9] ESA Publications Division, "Information/Documentation Management", ECSS-M-40, 13th June 2003

[10] ESA Publications Division, "Project Phasing and Planning", ECSS-M-30B, 13th June 2003

[11] ESA, "Selection Process for a Science Mission", http://sci.esa.int/science-e/www/object/index.cfm?fobjectid=33263

[12] ESA Board for Software Standardisation and Control, "Guide to applying ESA Software engineering standards to small software projects", ESA 1996

[13] Gaia Science Advisory Group, "GAIA Composition, Formation and Evolution of the Galaxy" ESA-SCI(2000)4

[14] Gilmore G., Perryman M., Lindegren L., et al., "GAIA: origin and evolution of the Milky Way" 1998, in R.D. Reasenberg (ed.), Astronomical Interferometry, SPIE Vol. 3350, p. 541

[15] Gray J & Compton M, "A Call to Arms", ACM Queue vol. 3, no. 3 - April 2005

[16] Handy Charles, "Understanding Organisations", 1993 Oxford Press.

[17] Kantor J, Axelrod T., Allsman R.,"Large Synoptic Survey Telescope Data Management Organization Charter" v1.1 (Private Communications)

[18] Lasker B.M., McLean B.J., Jenkner H., Lattanzi M.G., Spagna A., "Potential Applications of GSC-II for GAIA Operations" 1995 Proceedings of the workshop - "Future Possibilities for Astrometry in Space",p137 . ESA Sp-379.

[19] Lock D., "Project Management", 7th edition Gower 2000

[20] Maturana H. R., Varela, F. J. "Autopoiesis and Cognition: The realization of the Living". Boston Studies in the Philosophy of Science, Vol. 42. Reidel Doedrecht, 1980.

[21] Montmerle T., Turler M., "ISDC News Letter Number 5.", ISDC Website, March 2001 http://isdc.unige.ch/index.cgi?Newsletter+nothing+Newsletter/N05/index.html

[22] O'Mullane W. and Lindegren L., "An Object Oriented Framework for Gaia Data Processing". Baltic Astronomy March 1999.

[23] O'Mullane W. "Structures for DACC", GAIA-WOM-001C, Gaia Document Repository (private) August 2005.

[24] Stoughton C., "SLOAN DIGITAL SKY SURVEY: EARLY DATA RELEASE" THE ASTRONOMICAL JOURNAL, 123:485-548, 2002 January, http://www.journals.uchicago.edu/AJ/journal/issues/v123n1/201415/201415.html?erFrom=70003936241882830 11Guest

[25] Tigris Software, "Subversion version control" http://subversion.tigris.org/

[26] http://www.rssd.esa.int/SA/GAIA/docs/info_sheets/IN_astro_telescope.pdf

[27] http://www.rssd.esa.int/SA/GAIA/docs/info_sheets/IN_astro_focal_plane.pdf

[28] http://www.rssd.esa.int/SA/GAIA/docs/info_sheets/IN_measurement_principle.pdf

[29] http://www.rssd.esa.int/SA/GAIA/docs/info_sheets/IN_chromaticity.pdf

[30] http://www.rssd.esa.int/SA/GAIA/docs/info_sheets/IN_RV_objectives.pdf

[31] http://www.rssd.esa.int/SA/GAIA/docs/info_sheets/IN_spectro_focal_plane.pdf

[32] http://www.rssd.esa.int/SA/GAIA/docs/info_sheets/IN_Spectro_telescope.pdf

[33] http://www-1.ibm.com/servers/eserver/linux/power/marenostrum/about.html

Appendix 1. Answers from GSC/DSS Project

1 Introduction

This is the questionnaire used to guide interviews about projects for the study. A summary of the answers is provided below under each question.

2 Background

Q1. What is your name and position in the project?

Brian Mclean , Scientist

Q2. May I record this conversation?

Yes.

Q3. May I use your name in my final report?

Yes

Q4. Would you provide a brief summary of the project or salient reference ?

GAIA proceedings from Cambridge 1995, Summary of GSC and DSS, ES SP 379

3 Mission Costs

Q5. What was the overall budget estimate for the mission at the outset?

No good answer. GSCI was part of HST and this was expanded, there was external funding. Probably was formal budget but it is unknown to Brian who was not in management then.

110FTEs from project start GSC2 ~$2 million since 1998. GSCI was probably another 100FTEs.

Q6. What was the final/current overrun or under spend?

The best possible job was done with the available money

Q7. How much was earmarked for Software development?

Bulk of the cost (60-70%) was in personnel and they were doing software development.

Q8. Was there an over/under spend on software development, how much?

Not possible.

Q9. How much was earmarked for Hardware procurement?

The scanning machines were expensive but these were covered in the original HST operations costs. $1 million was spent on hardware for DSS1.

Q10. Ws there an over/under spend on Hardware, how much?

Possibly but it was all in the original HST ops.

4 Management

Management is a tricky topic but one for great interest to me, especially for science projects. To put this in perspective again we are interested here in the management of the data processing and storage teams not perhaps the building of the entire instrument and general project.

Q11. What size has the tem been over the lifetime of the project?

Fluctuated from 10 to 21 currently about 10 ramping down for last 5 years.

Q12. How many institutes are involved?

14 STSCI , 2 ESO, 5 at Torino. These are the 3 institutes in active work. Sponsored by 12 institutions.

Q13. Has a particular management style been consciously adopted by the project manager?

Barry (Laskar) ran everything. Barry was a leader – he did not adopt a style. He lead by example.

Q14. Are managers formally trained ?

Minimal. Some team building technical leadership.

Q15. Can a one page Organigram be provided?

Not in that form – could be done. Might have mind map (influence of the Author).

Q16. What would you change with the advantage of hindsight?

Not more control. But would like to make people more accountable to keep the schedule. Across the board reliability was a problem, went in with positive optimistic attitude , assumed others were similar. Did not always work like that.

Q17. What was the manpower/time estimate for your major software product ?

Tasks were generally estimated on a task basis.

Q18. What was the reality?

Even with experience the estimates were always underestimates even when doubled.

5 Software

Some of these of course could be answered with a yes or no but I am hoping for a little elaboration ☺

Q19. Are a set of software engineering standards used in the project (ISO,ECSS), is adherence checked ?

"We are not software engineers". There were no formal standards. Had own way of doing things. There were coding standards but no lifecycle. The entire project goal was known to everyone but it was not formalized down to lower details.

Q20. Were standards mandated by a funding agency?

No.

Q21. If standards were mandated state your opinion on their benefit to the project?

NA.

Q22. Is a particular Software development methodology used in the project or parts of it (OMT, Booch, Waterfall) ?

Rational Rose was used to do the DB design/Data Model for Objectivity and later for other small parts of the project.

Q23. If a methodology is used how was it selected?

Came from using objectivity. Objectivity probably because of influence from A. Szalay and SDSS.

Q24. Do you use a source code control system such as CVS?

For original processing software not. Code was copied to test and then to live. Source Safe was used for the C++ on windows when objectivity became the main system.

Q25. Do you use a release management system such as ClearCase?

None. Used MMS for a while (VMS system)

Q26. Do you use a problem tracking system?

No. Would now. Not even a single person in charge of problems – everyone pitched in.

Q27. Have you partnered with a major vendor for software production?

Not really. Worked with objectivity but no real on site help.

Q28. Have you been able to use COTS (Common Off The Shelf) components in you system? (Even Freeware)? Did it save money?

Objectivity. Unclear that it saved money – changed the way people worked and what they worked on. "No way to tell how much it saved us". GSC1 worked fine but no formal database but it would not scale like Objectivity did. But there was a big learning curve for Objectivity.

Q29. What is your main development language? (if you have one).

Fortran moving to C++ moving to C#. IDL is used throughout the project. Some Java for part of project.

Q30. How would you rate your processing in terms of difficulty , describe it a little?

Moderately complex algorithms, object detection and image calibration. Proper motion calculations etc were simple enough. Fairly complex to manage, so big, so much to manage. There were 8 Terabytes of images.

6 Hardware

Q31. What kind of hardware system do you have, monolithic mainframe/supercomputer or cluster/distributed system or something else entirely?

Evolved. Started with VMS clusters, servers and workstations which also did processing. VMS servers still running. Then went to windows. Now one big windows box. Still processing images on VMS. Data loaded in DB but some images need to be reprocessed e.g. in galactic plane where object density is high.

Q32. Approximately how much processing power have you got?

No idea. Not even a retrospective calculation. A notion of time per image was known and more hardware was purchased to make it faster.

Q33. How much disk space?

Now have 25 Terabytes spinning. Interesting that this is 3 times raw data. Raw images one third, scratch one third, one third for result databases.

Q34. How much disk and processing power was estimated for in the beginning of the project (if one was made)?

Q35. Do you use a tape archive? If so is it still cost effective?

Backup is to tape. But not reliable enough. Guaranteed an error in 40GB backups. Moving all to spinning RAID array.

Q36. Have you partnered with a major hardware vendor? Was it successful? Did you ever feel locked in?

DEC in the beginning , moved to wintel. No feeling of lock in. Objectivity drove wintel decision also. VMS for image processing, Database on windows. Mainly DELL (institute decision) but not exclusive. Obviously no lock in since a move was made.

Appendix 2. Answers from SDSS Project

1 Introduction

This is the questionnaire used to guide interviews about projects for the study. A summary of the answers is provided below under each question.

2 Background

Q1. What is your name and position in the project?

Bill Boroski, Project Manager for Sloan Sky Server

Q2. May I record this conversation?

Yes.

Q3. May I use your name in my final report?

Yes

Q4. Would you provide a brief summary of the project or salie nt reference ?

Sloan Digital Sky Survey is a project creating a digital map of ¼ of the universe doing both imaging and spectroscopy. Currently seeking funding for 3 more years which would make an 8 year survey. When done it will be close to 8000 square degrees of imaging data and 1 million redshifts of galaxies quasars etc. Data will be made available to the public from one main site and mirror sites all over the world.

3 Mission Costs

Q5. What was the overall budget estimate for the mission at the outset?

~$25 million to do a five year survey

Q6. What was the final/current overrun or under spend?

Final project for the 5 year survey is $85 million. The project was vastly under scoped in terms of work required and hardware and procurements.

Budget for 3 more years is $15million. Current running costs are $5.5 million per year.

Q7. How much was earmarked for Software development?

Not available.

Q8. Was there an over/under spend on software development, how much?

Probably also vastly underestimated. All development was supposed to be done before operations but it went on well in tot the first year or year and half of operations. There is still development work on calibration software and databases.

Q9. How much was earmarked for Hardware procurement?

No breakdown at the moment.

Q10. Was there an over/under spend on Hardware, how much?

Underestimated both for the mountain top and processing.

4 Management

Management is a tricky topic but one for great interest to me, especially for science projects. To put this in perspective again we are interested here in the management of the data processing and storage teams not perhaps the building of the entire instrument and general project.

Q11. What size has the team been over the lifetime of the project?

More familiar with last 1.5 construction up to now say 1997. We had a 4 person management committee which forms the core of the management team. Then five level one managers who oversee systems like the observing systems (everything that supports mountaintop operations), data processing and distribution, observatory itself, survey coordination and a business manager. Core Management team is around 9 or10 people.

Q12. How many institutes are involved?

14 institutions involved now. 7 have people very actively involved in the infrastructure associated with the project.

(Fermi, Hopkins, Princeton, Chicago, Naval observatory, University of Washington, New Mexico, Chicago state – may be missing one.)

Others provide funding or get involved in a project like calibration task force.

Q13. Has a particular management style been consciously adopted by the project manager?

Collaborative effort. Try to manage in a collegial manner but still top down – core team sets direction. Management delegated down to the different levels. E.g. I look after observing systems with Jim Gunn, as long as everything is ok we are left to ourselves.

Q14. Are managers formally trained?

One formally trained manager – that's me. Mike Evans has formal training also.

Q15. Can a one page Organigram be provided?

To be provided. ARC (Astrophysical Research Council), a group of universities, is the legal entity running the project. A sub body called the advisory council which advises ARC and delegates the day to day ops to Rich Kron the Director. Director created the management committee to help him. The committee is the director, the project scientist, the project manager and the project spokesperson. We deal with collaboration issues (the spokesperson), the project manager deals with all cost and schedules and day to day operations, the project scientist deals with science issues. The director oversees all of this.

Q16. What would you change with the advantage of hindsight?

There have been a lot of occasions were we have not been good at clearly defining the roles and responsibilities of different individuals in key positions, giving them the authority for doing that job and holding them accountable for doing that job. That led to some confusion.

Problems have occurred where people have said " you gave me this position and title, now let me do my job". But you can not tell people "just do it" there is lot of negotiating and coaxing. Getting people to do what you want is challenging, some people can deal with it better than others. So that's an issue holding people accountable for what they have done.

Another issues goes back to the scientists. A lot of people were used to dealing with small science projects within their own organisation or lab and Sloan is a big science project. Have to deal with a lot of transparency, formal procedures. For example "Why do version control?". Now on the mountain a system is in place for delivery and testing of code before it goes in production the old way in the middle of a run the software version may change and we would not know what happened. When we tried to implement formal procedures people balked at it saying it was bureaucracy.

Another issue is the "unevenness " of the skill level in the team. So some people would say (procedures) are fine for everyone else but do not apply to me.

In hindsight if the policies, procedures version control etc. were in place up front and you were disciplined about following them the operation would run smoother. Some people would be unhappy initially. Management need to behind it.

Q17. What was the manpower/time estimate for your major software product?

Unsure

Q18. What was the reality?

It overran. 4 institutes originally . Work was done on an ad-hoc volunteer basis – everyone went off thinking that it would all come together but it did not of course. No one was to blame, just no one ever did this before and it was a hard job.

5 Software

Some of these of course could be answered with a yes or no but I am hoping for a little elaboration ☺

Q19. Are a set of software engineering standards used in the project (ISO,ECSS), is adherence checked?

No formal standards for software. But it might have helped. Early on there were requirements and statement about platforms and languages. Don Petravick and the data acquisition team wanted clearly defined requirements and they were going to build to those requirements. They had a very rigorous approach but I do not know if they used formal standards. In contrast Robert Lupton and those at Princeton, they had an idea working with Jim Gunn of what needed to be done and they worked towards that. But they never signed off on requirements so they could say they were finished.

Q20. Were standards mandated by a funding agency?

No.

Q21. If standards were mandated state your opinion on their benefit to the project?

Q22. Is a particular Software development methodology used in the project or parts of it (OMT, Booch, Waterfall)?

None – everyone has their own way. When one takes over another's code a lot of rewritten is done. Not everybody but a lot of people – seen a lot of code tossed. People argue I understand my code better – quicker to rewrite. There are some parts no one will touch

Q23. If a methodology is used how was it selected?

Q24. Do you use a source code control system such as CVS?

CVS. Widely used. Some problems getting buy in. It's a very good idea.

Q25. Do you use a release management system such as ClearCase?

Nothing formal. Wish we did, wish I knew more about it. There is a system whereby developers tag a module for release – another person checks this out and builds it, another tests it, before it is declared for release.

Q26. Do you use a problem tracking system?

Gnats. It works. For Problem Reports (PR) and managing Change Requests (CR). On the mountain there is a formal system to look at PRs and CRs and review them and decide which ones to deal with and which to have work around for.

A big part of management is that it is easy to have good ideas its really hard to execute. Robert is often on my back that I am not keeping on top of the PRs. We should not have open PRS we should not have critical high PRs. It would be nice to have someone responsible to chase them down. System is only as good as how well you use it.

Q27. Have you partnered with a major vendor for software production?

no

Q28. Have you been able to use COTS (Common Off The Shelf) components in you system? (Even Freeware)? Did it save money?

Yes. ImageMagick, TCL, SQLServer, MySql freeware. Certainly saved money yes.

Q29. What is your main development language? (if you have one).

C but all kinds of other things

Q30. How would you rate your processing in terms of difficulty, describe it a little?

Very Challenging. In terms in data collection, processing and Management of the data. Huge amount of effort went into making the processing a factory, automated and modular as possible. Chris Stoughton gets credit for that.

6 Hardware

Q31. What kind of hardware system do you have, monolithic mainframe/supercomputer or cluster/distributed system or something else entirely?

Distributed heterogeneous.

Q32. Approximately how much processing power have you got?

No idea.

Q33. How much disk space?

Will be provided – on the order of 40Tb for finished data then there is scratch space and desktops and on the mountain (maybe 2 Tb up there).

Going to have 40tb for 3.6Tb of data (that's the SQLServer database).

Q34. How much disk and processing power was estimated for in the beginning of the project (if one was made)?

Q35. Do you use a tape archive? If so is it still cost effective?

Yes ENSTORE. DLT Tapes are sent (FedEx) from the mountain and put in ENSTORE to get the data – this is more cost effective than streaming the data. This is also used for backups. 9 DLTs for an image. Write to 2 sets of tapes. One set shipped one stays. Once Per year other set put in cold storage. Have actually retrieved some occasionally and they have worked.

For upgrades talking about writing to hot swappable IDE drives. But IDE disks are now cheaper. Princeton have a special shipping case

Q36. Have you partnered with a major hardware vendor? Was it successful? Did you ever feel locked in?

No. Have occasionally gotten some hardware from some vendors.

Q37. Anything else to add?

Requirements are very important, getting scientist to agree to the requirements up front is very hard. Its one of the big challenges of this business. If you don't have them you don't know when you are done and it is very hard to manage with a fixed budget.

Several early operations reviews consistently mentioned, Developers have a different mindset than operations people, a lot of times people can not make the transition. Some people are of an ops mindset they say we are done it meets requirements we are running with it, while a developer will say 'yes' but I know I can make it better. For a project like Sloan, an industrial strength science project, at some point you just want to shoot the developers and say we are done. It has been a challenge; some reviewers suggested looking at staff and possible replacing some developers with operations people. It is not easy to go from construction to ops with the same group of people.

There is a reason people are where they are. There is a reason that people in this project are in academic institutions because they like that culture and that freedom do not want the rigor of

the corporate structure. Try to get the rules and procedures in place first, early and them enforce them rigidly. I have had to change people's passwords – they would not follow protocol and I changed the password. Bad thing to do. Annoys some people, gains some support from others. But sends a clear message that the rules are important.

If everyone knows there is a problem that needs to be taken care of it needs to be taken care of. It causes bad moral all round and management are not seen as dealing with it. Early in Sloan there were individuals in the way, a source of problems, everyone knew it, but management, the people who could have done something about it, did nothing. When there was a management change and those people were dealt with there was a vast improvement in moral. Don't let things fester. If something needs to be done do it, but it takes guts.

Appendix 3. Answers from LSST

1 Introduction

This is the questionnaire used to guide interviews about projects for the study. A summary of the answers is provided below under each question.

2 Background

Q1. What is your name and position in the project?

Jeffrey Kantor, Project Manager for Large Synoptic Survey Telescope

Q2. May I record this conversation?

Yes.

Q3. May I use your name in my final report?

Yes

Q4. Would you provide a brief summary of the project or salient reference?

Project is to create a very wide aperture deep field telescope, it is as yet unclear if this will be in the northern or southern hemisphere. Then to do a continuous survey over 10 years in 5 filter bands of the entire half sky. Each image will be about 3.5 GigaPixels, shooting for 1% photometry and .2 arcsecond astrometry. We hope to support a number of different missions~: weak lensing science, galactic structure studies, solar system inventories, fast moving objects. My part of it is as each of those images comes out of the focal plane, there is one every 15 seconds or so, I have to do all the processing, reduction storage, curation of the storage, make it available for public science.

3 Mission Costs

Q5. What was the overall budget estimate for the mission at the outset?

There will be 3 phases.

R&D , studies, proof of concept etc. will be ~$14 million government funding plus $10 - $14 million of private funding.

Construction, build the telescope set up the data canters, configure it all commission it and so on. ~$270 million.

Operations, initial survey period of 10 years to get really good catalogues, and images from that ~$20 million per year, about half for data management. 3 Petabytes per year.

Q6. What was the final/current overrun or under spend?

Can't overrun, R&D stops when the money is gone.

Q7. How much was earmarked for Software development?

About $60 million during construction phase for data management. Hardware, software, everything once the bits leave the camera.

Q8. Was there an over/under spend on software development, how much?

Not in that phase yet.

Q9. How much was earmarked for Hardware procurement?

60% of the 60 million is going to be for software development and 40% for hardware. Not all the hardware will be purchased at the beginning of the mission, since it will get cheaper.

Q10. Was there an over/under spend on Hardware, how much?

Not in that phase yet.

4 Management

Management is a tricky topic but one for great interest to me, especially for science projects. To put this in perspective again we are interested here in the management of the data processing and storage teams not perhaps the building of the entire instrument and general project.

Q11. What size has the team been over the lifetime of the project?

Just assembled leadership team for Data Management. Jeff, Tim Axelrod (Project Scientist) six months ago. Permanent staff will probably peak at 15-16 people. Also a large number of volunteers currently involved in R&D phase.

Q12. How many institutes are involved?

Now 12 members of the consortium (in America), organized into 3 research teams. There are 6 working groups to define requirements.

Q13. Has a particular management style been consciously adopted by the project manager?

Democratic manager. Like to gather broad range of input, vet that through a decision process. Pretty strict once decision is made not to lightly or arbitrarily overturn a decision. Need to keep things moving.

Q14. Are managers formally trained ?

Yes software developer and IT manager for many years, with lots of formal training (technical, management, quality assurance).

Q15. Can a one page Organigram be provided?

Formal org chart is future state diagram. Currently 3 research teams and they have a temporary structure. The 16 people structure is the target and has been documented.

Q16. What would you change with the advantage of hindsight?

This is the beginning, don't have any hindsight yet.

Q17. What was the manpower/time estimate for your major software product?

16 people over 45 years. I use a formal estimating methodology based on COCOMO, FPA, and SLIM/QSM. The estimates I ran actually suggest it would be possible to do this in a shorter period of time with more people. But because of the research nature of some of the algorithms we are not sure how best to do them yet. I think that we will find is time will be stretched and we will be able to do it with a lower number of people over a longer period.

Q18. What was the reality?

Not in this phase yet.

5 Software

Some of these of course could be answered with a yes or no but I am hoping for a little elaboration ☺

Q19. Are a set of software engineering standards used in the project (ISO,ECSS), is adherence checked ?

Using a UML based specification process called the ICONIX process. It is sort of half way between the very formal heavy specification process and the very agile eXtreme programming that has almost no specification processes. About two thirds of the way over to the extreme side. You can't go all the way over there because then you are just hacking, but you can still stay agile and do some specifications.

I am a big believer in models and prototypes, I am not a big believer in documents except for user documentation. We work on the models and when we have to we produce the documents from the models (UML modelling, Use cases etc).

Q20. Were standards mandated by a funding agency?

We have some project management standards we expect to be mandated by the Department of Energy pert charts, earned value reporting, etc.. Even NSF for the large MRE (Major Research Facilities) type projects, is starting to require more formal project documentation

Q21. If standards were mandated state your opinion on their benefit to the project?

Mixed opinions about that. If done properly they are useful reporting mechanism, but sometimes they are a heavy burden just preparing documentation. The larger the project the more you need it.

Q22. Is a particular Software development methodology used in the project or parts of it (OMT, Booch, Waterfall)?

The Iconix methodology is more a technical methodology than a project management methodology. Historically, OMT, Booch and Objectory were 3 predecessor methodologies to the Unified process, a predecessor to the Rational unified process. At the same time Iconix were developing their methodology and kept it a little lighter. The 3 gurus (Booch, Jacobsen, Raumbaugh) did a good thing trying and stop the methodology wars, but they (Rational) wrote an encyclopaedia and no one understands how to apply the encyclopaedia.

Q23. If a methodology is used how was it selected?

Experience over many years and projects of Jeff and Tim.

Q24. Do you use a source code control system such as CVS?

Using CVS, least common denominator. Looking at Subversion (said revision but later corrected).

Q25. Do you use a release management system such as ClearCase?

Looking for something a little lighter weight.

Q26. Do you use a problem tracking system?

Do have an issue and risk tracking system. Microsoft project Server which has a component for this but not for defect tracking. We will have a software defect tracking system also, when we are producing significant amounts of software.

Q27. Have you partnered with a major vendor for software production?

Not currently.

Q28. Have you been able to use COTS (Common Off The Shelf) components in you system? (Even Freeware)? Did it save money?

We are still testing DBMSs. We have 3 layers: application layer, middleware layer, and infrastructure layer. The infrastructure is hardware and system software we anticipate that will be 99% off the shelf the only area where there might be something special purpose is in the acquisition interface to the camera. Middleware we anticipate using a lot of off the shelf software, probably Condor and a lot of the Grid tools, probably MPI (Message Passing Interface) for some of the pipelines. We will definitely be using some form of database management system for portions at least of our catalogue (it may be commercial or open source). In the application layer that's where most of the custom work will be, specific algorithms etc. Certain parts of the problem lend themselves to files systems and some to DBMS. Do have some current issues with DBMS performance but we are researching query parallelization and rapid ingestion to address that.

Q29. What is your main development language? (if you have one).

Baseline is C++ and Python, may extend to Java for less performance critical or web-centric work.

Q30. How would you rate your processing in terms of difficulty, describe it a little?

The data rate is unprecedented, a DVD worth of data every 15 seconds. We have transient alerting requirements which are sub a minute. We have to correct and register the images, we have to photometrically and astrometrically calibrate them, we have to classify, detect and alert and we've got to do that in less than a minute. We also have to provide feedback from a quality control standpoint back to the telescope in less than a minute. If an image is not looking good it needs to be done again or adjustments need to be made, we have to point out 'here is where its out of whack as far as we can tell'. This will be done at the mountain base.

Accumulating the data at that rate becomes on the order of 2-3 Petabytes per year. Being able to efficiently query that, search it, is another challenge.

A longer processing will also be done less frequently, images will also be stacked. This will be done at the archive centre and fed back to the mountain base for future subtractions and so on.

6 Hardware

Q31. What kind of hardware system do you have, monolithic mainframe/supercomputer or cluster/distributed system or something else entirely?

We have 3 kinds of computing centre each with more than one subsystem each with its own hardware configuration and architecture.

On the mountain base we have the data acquisition plus a pipeline server and storage sufficient to support the pipeline server and buffer 2-5 days of data.

We anticipate having a 2.4 to 4 gigabit link to the archive centre. We are going to ship the raw data, even though its already been processed. We believe the long haul link is a cost-driven limitation. We are monitoring the network availability, we feel confident by the time we go operational that 2.4 will be available and maybe as much as 4. So we are architecting to that capacity. That means we do not want to ship both the raw data and the processed data over that link. So we will send the raw data and process it all over again at the archive because computers will be a lot cheaper than long haul bandwidth.

The archive centre also has Data Access Servers, this is where the VO would already come in for event alerting and data access.

Then we have pure data centres which are replicated sets or subsets of the data for general availability. Those only have data servers. We have a tiered access model for getting at the data to manage performance.

Q32. Approximately how much processing power have you got?

We figure aggregate between the mountain base and one archive centre with pipeline server and data access server we figure we need ~75 TeraFlops.

There will be other centres to optimise community data access.

Q33. How much disk space?

Uncompressed 3 Petabytes a year so probably 2 times that amount (not sure). Cant imagine more than 3.

Q34. How much disk and processing power was estimated for in the beginning of the project (if one was made)?

Disk is getting cheaper than tape. Disks may not be as reliable over the long term there is a real trade off. Working with NCSA which has huge disk and tape farms, Livermore and

Brookhaven have huge disk and tape farms. They are all looking at this. We are going to let the big centres tell us where to go with this.

Part of the premise is any one part of the survey is as interesting as any other part – an argument for keeping it all on disk.

Q35. Do you use a tape archive? If so is it still cost effective?

See previous question.

Q36. Have you partnered with a major hardware vendor? Was it successful? Did you ever feel locked in?

Not so far. Looking at IBM blue gene and Cell (IBM, Sony) architectures, others as well.

Q37. Anything else to add?

Volunteers are definitely a two edged sword. You get some value but you really have to work at it!

Appendix 4. Answers from CDS

1 Introduction

This is the questionnaire used to guide interviews about projects for the study. A summary of the answers is provided below under each question.

2 Background

Q1. What is your name and position in the Facility?

Francoise Genova – Director of the Centre de Données astronomiques de Strasbourg (CDS)

Q2. May I record this conversation?

Yes.

Q3. May I use your name in my final report?

Yes, okay.

Q4. Would you provide a brief summary of the facility or salient reference?

CDS is a data centre which was founded in 1972 by the Institut National d'Astronomy et de Géophysique which takes care of ground based facilities in France which is now INSU. CDS has several roles, one is providing reference services, added value services to the astronomical community. From the very beginning it was put in an observatory so there was scientific expertise and to keep the focus on serving astronomers and not building technical tools. We have a lot of experience building reference tools and standards, so we are also leading the French effort in the virtual observatory (VO). In a sense we were precursors of the VO at the international level

3 Costs

Q5. What is the budget of your facility?

It is difficult to compute the cost completely as we rely on government positions. So I do not get a budget rather a number of staff. Most of the astronomers and technical staff are in government positions.

Q6. Is there a typical overrun or under spend on projects?

We check the project status very closely to see if they are taking too much time. If a project is taking too much effort I need to take care of the global balance. If something is more difficult

than expected we may drop an action (Requirement). Some projects are on specific budget but usually, on European projects, we try to define what we will properly do and do it properly.

Q7. How much was earmarked for Software development?

(wil) You have mostly software at CDS? We have quite a bit of hardware. Hardware is managed by the sys admin and his aide at the observatory. The main cost is people, not only software engineers, also astronomers and a lot of people who are trained as librarians who build the database contents Simbad or Vizier contents. So you do not know them but there are many, there are more documentalists than software engineers working on the project. Many of the astronomers work on the content and not working directly on software development.

Q8. Is there an over/under spend on software development, how much?

Not relevant.

Q9. How much is earmarked for Hardware procurement?

Small part of overall cost.

Q10. Is there an over/under spend on Hardware, how much?

(will)You have what you have ? Yes and we adjust to it.

4 Management

Management is a tricky topic but one for great interest to me, especially for science projects. To put this in perspective again we are interested here in the management of the data processing and storage teams not perhaps the building of the entire instrument and general project.

Q11. What size has the team been over the lifetime of the facility?

When CDS began it was only a few individuals and for fifteen years there was only one software engineer. Then it grew progressively, with people coming and going. We are now between 25 and 30 but not all full time and some not in Strasbourg.

Q12. How many institutes are involved or is it all in house?

Most are in Strasbourg. We have a few librarians in two institutes in Paris. There are two Astronomers with content expertise are in other French towns. So five institutes in all.

Q13. Has a particular management style been consciously adopted by the management ?

No particular style, no standard definition. We try to have meetings with a broad range of profiles, software engineers, astronomers together and discuss the status of different projects and the global strategy. Consensus building and taking into account the different points of view on the direction of astronomy which must be taken care of in databases, what is the evolution of technology which means we need software engineers who do real technical work. We need also to see the policies of the agencies and understand how to respond. Also what are the possible collaborations.

Q14. Are managers formally trained?

no

Q15. Can a one page Organigram be provided?

I am working on it. I have to take in to account the cross project dependencies and transverse expertise. I have to find a matrix organigram. Three dimensional perhaps.

Q16. What would you change with the advantage of hindsight?

Not always happy but it is hard when one always tries to perform the best. One must also consider when you do something differently it is possible to see which other things may have also come out differently – overall it is difficult to say if a change would have made anything better in the end.

Q17. What was the manpower/time estimate for one of your major software products? What was the reality?

Irrelevant.

5 Software

Some of these of course could be answered with a yes or no but I am hoping for a little elaboration ☺

Q18. Are a set of software engineering standards used in the faculty (ISO,ECSS), is adherence checked?

No particular standards.

Q19. State your opinion on their benefit to the facility?

Q20. Is a particular Software development methodology used in the facility or parts of it (OMT, Booch, Waterfall, Unified, Iconix)?

No. Some people use tools – it's an individual basis.

Q21. If a methodology is used how was it selected?

Q22. Do you use a source code control system such as CVS?

We have begun to use it. It is useful.

Q23. Do you use a release management system such as ClearCase?

Q24. Do you use a problem tracking system?

No formal problem tracking.

Q25. Have you partnered with a major vendor for software production?

Q26. Have you been able to use COTS (Common Off The Shelf) components in you system (particularly DBMS)? (Even Freeware)? Did it save money?

We use several databases. We look at requirements and choose. We have relational, OO and in house database systems. For long term maintenance we are trying to move in house things to Postgress (Simbad is being ported). – just a remark here: this does not mean we will use Postgress for all other CDS database needs since we do case by case requirement study -

Q27. What is your main development language? (if you have one).

It used to be C,C++ but many people are using Java now.

Q28. How would you rate your processing in terms of difficulty, describe it a little?

We have to read the journals and get the Simbad information. It is difficult but relies on people. Effort has been put in to define procedures. We are revisiting it.

6 Hardware

Q29. What kind of hardware system do you have, monolithic mainframe/supercomputer or cluster/distributed system or something else entirely?

No supercomputer. Some Unix and more and more PCs. Operational machines are Linux. – also PC clusters

Q30. Approximately how much processing power have you got?

Q31. How much disk space?

More than 1Tb , 5 or 6 terabytes on several machines. Nothing special. The first Simbad was on a mainframe. It went from IBM to another centre and to Unix – it has moved many times.

Q32. Do you use a tape archive? If so is it still cost effective?

We have backups. But we are trying to have a full copy on disk in another building. Tapes take a long time to rebuild a system. Cost was an issue.

Q33. Have you partnered with a major hardware vendor? Was it successful? Did you ever feel locked in?

Q34. Anything else to add on any topic?

Adjust the management to the project. Understand the conditions of the project, the people, the partners and the goals (what you want to do) and if you have many organizations what are their own constraints. Define roles and goals for each organization and get them to agree on them. Adjust the management and management style to the goals of the project, this is just common sense. But it is not always possible, if you are in a highly organized structure you are not free to choose your organization principles.

Appendix 5. Answers from Integral Project

1 Introduction

This is the questionnaire used to guide interviews about projects for the study. A summary of the answers is provided below under each question.

2 Background

Q1. What is your name and position in the project?

Lars Hansson, Integral Science Operations manager.

Q2. May I record this conversation?

Yes.

Q3. May I use your name in my final report?

Absolutely

Q4. Would you provide a brief summary of the project or salient reference?

I am responsible for one half of the science ground segment. The ground segment normally consists of an uplink part and a downlink part. The downlink part, data processing and distribution to the community is done by a PI consortium at the Integral Science Data Centre (ISDC) under the Observatory of Geneva. So ESA is doing the uplink part, I have been responsible since we started the development and am now also managing the operations. The software was done in house at ESTEC. ISDC also came through me for interactions with ESA. Now there is a triumvirate, myself, Roland Walter (ISDC) and the SOM (Spacecraft operations Manager) in MOC (Mission Operations Centre in Germany). ISOC is also maintaining a copy of the Scientific archive. The importance of Scientific Archives are much more pronounced now in the Agency. In agreement with ISDC ESA is building a user interface corresponding to the corporate look and feel already used in other ESA mission archives. The public part of the ISOC archive will also be made available to the science community. This version is now undergoing beta testing before being made available to the public. The development is made by the archive group at ESAC. ISDC is using a user interface developed by HEASARC in the US.

3 Mission Costs

Q5. What was the overall budget estimate for the mission at the outset?

This is a midsize mission overall 400 to 500 million Euro. That includes the ESA part and the contributions made to the instruments. ISDC are funded independently but are well supported by the agency. We have put in quite some support to ISDC, for example for the archive development, QA, testing and administrative support. ESA has put approximately 20 man years of effort into ISDC.

Q6. What was the final/current overrun or under spend?

The overall mission was well under control. It did not even use up all the margins. We invested more in the instruments than originally budgeted.

Q7. How much was earmarked for Software development?

ISOC was about 4-5 people for 6-7 years about 40 man years. In addition around 5 man years have been spent up to now on the ISOC science archive. ISDC were 25-30 people for around 8 years so around 200 man years.

Q8. Was there an over/under spend on software development, how much?

ISOC was always under control and within budget. For ISDC I did not have insight but things are often a little different with institutes.

Q9. How much was earmarked for Hardware procurement?

250-300 KEuros at the moment. But it is not sufficient – ISDC reprocessed and the volume has grown.

Q10. Was there an over/under spend on Hardware, how much?

Saved by falling prices of hardware. We need much more disk than originally anticipated. When we sized the archive we did not anticipate the expansion in size of the new processed data – many new products which were not foreseen. But its ok for a year of data storage its only 15K.

4 Management

Q11. What size has the team been over the lifetime of the project?

4-6 ISOC, ISDC 25-30.

Q12. How many institutes are involved?

ISOC only ESA all in house. ISDC has many institutes also including non EU institutes like Poland and Czechoslovakia – around 10 institutes.

Q13. Has a particular management style been consciously adopted by the project manager?

ISDC is a university type non managed environment. ISOC is highly managed in a top down manner. Christine Brenol was in charge of the development and I took it over when she left.

Q14. Are managers formally trained?

Not particularly – Christine was sent on a course. I have more experience than formal training – some courses.

Q15. Can a one page Organigram be provided?

Would have to look in on the web sites. We just moved and cleaned up so it may be difficult to look back in history. There may be something on the website. We kept our docs in CVS and put them on a website – no LiveLink or document management system was used.

Q16. What would you change with the advantage of hindsight?

Well one thing since you are here[23]. We started off in a very OO way. But somewhere along the way we need to break that. It was good for gathering requirements and making the architecture. But from then on to maintain it was too much effort really – I would have spent half the development effort on maintaining the ADD. Java is quite reasonable in self-documenting. Half way through we exchanged the database and the proposal handling system. The solution we had was too thin the database was not deep enough. I got a new contractor in who developed a new system in fairly short time. That was around 2 years before launch. I keep compatibility – you could switch between old and new. I am very happy we did that. It is doubtful we would be able to import parameters from MOC in the old system. With help of the technical directorate we started a little project to look at this. Oracle was integrated into ISOC more completely. The new contractor also introduced Jbuilder and Junit. This is very good. It was an interesting process to train the team – they were a little reluctant. The new proposal handling was written using Junit tests. Now it is being retrofitted to the other systems. Java choice was a good one.

[23] William O'Mullane set up the Rose system for generation of SRD and ADDs for ISOC as well as the requirement matrices.

Q17. What was the manpower/time estimate for your major software product?

ISOC worked out as expected. You don't get the requirements you find out the hard way. I stayed on the manpower level. We also had a delay of one year which helped. ISDC were not that critical for launch.

Q18. What was the reality?

5 Software

Some of these of course could be answered with a yes or no but I am hoping for a little elaboration ☺

Q19. Are a set of software engineering standards used in the project (ISO,ECSS), is adherence checked ?

We used PSS-05 – we did not need to switch. The answer to the question is yes. We always had QA support. One of there duty is to ensure adherence to set standards.

Q20. Were standards mandated by a funding agency?

Q21. If standards were mandated state your opinion on their benefit to the project?

They were useful. I still maintain configuration control – the scientist do not like it but they begin to appreciate it.

Q22. Is a particular Software development methodology used in the project or parts of it (OMT, Booch, Waterfall)?

We started with a unified process but we had to change it a bit along the line. It is a question of individual experience. The ADD is no longer maintained and the all changes go through the CCB. I have always kept QA involved in the decision process. MOC side did not have QA involved and got a lot criticism during Launch Readiness Review. I had good test documentation acceptance test document etc. with links to SPRs.

ISDC had QA involvement also – ESA paid a good QA person to go to ISDC and set up QA for them.

Q23. If a methodology is used how was it selected?

Experience of team/team leader.

Q24. Do you use a source code control system such as CVS?

CVS was used

Q25. Do you use a release management system such as ClearCase?

Tagging etc. Nothing special.

Q26. Do you use a problem tracking system?

Taken from ISO and still in use. That was an in house system. Christophe has another better system. I supplement mine with a spreadsheet – Christophe's system does this automatically.

Q27. Have you partnered with a major vendor for software production?

ISDC partnered with CERN, ROOT

Q28. Have you been able to use COTS (Common Off The Shelf) components in you system? (Even Freeware)? Did it save money?

ROOT at ISDC. Java, Oracle, Jbuilder , (Rose initially) at ISOC. I would say it saved money and I have not seen many problems – number of SPRs is going down.

ISDC have also got a stable system – it depends on getting the right guys aboard.

Q29. What is your main development language? (if you have one).

ISOC – JAVA. ISDC – C++

Q30. How would you rate your processing in terms of difficulty, describe it a little?

The main parts of the ISOC core system, the proposal handling and the associated database and the mission planning are complex functions. A main complexity in ISOC was talking to flight dynamics software which was only available in binary form and written in FORTRAN. We have to kept a Solaris machine for that software still.

The science processing at ISDC is complex. In the Gamma domain the processing is very difficult compared to say XMM, it took much longer in Integral to get calibrations etc. This is genuinely to do with the difficulty of processing the Gamma Rays. The understanding of the instrument is now much better and there is a noticeable difference over the last half year.

6 Hardware

Q31. What kind of hardware system do you have, monolithic mainframe/supercomputer or cluster/distributed system or something else entirely?

Solaris because of Flight dynamics. Now we Linux PC for cost and performance. Now SUN have come back with something. ISDC has at least originally used a lot of SUN / Solaris computers. Whether they have moved on to replace with LINUX computers in not known.

Q32. Approximately how much processing power have you got?

Processing power has never been a critical issue since the proposal generation is done locally at the proposer's local computers. Then the proposals hare sent to ISOC for further processing and ingestion into the proposal data base.

Q33. How much disk space?

Around 6Tb +additional 5Tb just ordered.

Q34. How much disk and processing power was estimated for in the beginning of the project (if one was made)?

We estimated 6 at mission start – but now it will be more like 12Tb. This is probably not enough if considering a four year extension of the mission.

Q35. Do you use a tape archive? If so is it still cost effective?

No tapes.

Q36. Have you partnered with a major hardware vendor? Was it successful? Did you ever feel locked in?

no

Q37. Anything else to add?

It makes a big difference what type of guys you have on board. Careful selection of the team is important – I also see that the OSS is a very complicated beast but I have a much more automated system than XMM because I have automated a lot of things which I think they do manually. I had an expert in scheduling – XMM contracted that out. Get people with experience in the field you are working in.

Appendix 6. Answers from the Planck project.

1 Introduction

This is the questionnaire used to guide interviews about projects for the study. A summary of the answers is provided below under each question.

2 Background

Q1. What is your name and position in the project?

Jan Tauber Project Scientist

Q2. May I record this conversation?

Yes.

Q3. May I use your name in my final report?

Yes

Q4. Would you provide a brief summary of the project or salient reference?

It's a cosmology mission and its trying to make an image of the fluctuations in the Cosmic Microwave Background (CMB), the highlight will be if we can measure the polarized component of that.

3 Mission Costs

Q5. What was the overall budget estimate for the mission at the outset?

When we were selected (not with Herschel) we had a mission envelope around 350Million from ESA. The institute part was quite large also, today the two instruments are about 250Million. So a total of over 500million for Planck. Of course now we are with Herschel and the two can not be separated – for the two together the ESA part is about 1.1 to 1.2 billion and the whole thing probably close to 2 billion.

Q6. What was the final/current overrun or under spend?

This is controversial. Some people say yes it is costing more some say not. They put the two missions together essentially to save money. If you compare to the very optimistic and idealised estimates at the time of merger you could say we are overrun but if you go back to the original costs we have not overrun. We are having difficulties with the (ESA) science budget in general and this is the biggest program around so it always causes controversy.

Q7. How much was earmarked for Software development?

When we started out we tried to scale from previous missions. We tried to scale from previous missions. We had two data centres. The closest missions were Hipparcos and ISO. The estimates which came from the PI institutes was that each of the two DPC would need about 300 to 350 man years worth of effort. We are cost limited.

Q8. Was there an over/under spend on software development, how much?

Remains to be seen

Q9. How much was earmarked for Hardware procurement?

Difficult to say. Each of the two DPCs have some amount earmarked. But not high end super computers. Something on the order of a few million at most. This is still being discussed. Now the institutes are looking for post flight funding.

Q10. Was there an over/under spend on Hardware, how much?

4 Management

Q11. What size has the team been over the lifetime of the project?

Sticking with the science processing. This is tricky to estimate. The DPC rely a lot on academic personnel. These are often partial – very little of each individuals time is spent on the project. As we go toward launch things are crystallising to core teams – in both DPC we see a core of 20-30 people who spend at least 50% of their time on the project. There is maybe a smaller core team of less ten possibly five to eight people who spend 100% of their time on the project. So during operations we will have a core team of about 10 people and a floating team of about 10 more close to the core and many more around that (at each DPC). How to run operations is under discussion at the moment. They are talking about a team of people close to the mission of about fifty people at each consortium. That covers operations, processing and infrastructure.

Each of the consortia are much larger they have about 300 scientists. ESA play no role in data processing.

Q12. How many institutes are involved?

About 30 institutes with significant involvement but about 50 in total. These are categorised in the two DPC – a few are common.

Q13. Has a particular management style been consciously adopted by the project manager?

There is no overall style and each consortium is different. The Italian consortium is quite hierarchical with a fixed core and looser connection to scientific groups. The French consortium have a very loose management – the French academics tend to be more indivualist and do things out of good will rather than through direction. They also relay on a British group which makes things difficult to manage.

In the one case you have something more structured but with less drive and in the other case something less structured but with more drive, a more enthusiastic team.

Q14. Are managers formally trained?

No. We are becoming more focused. It is hard to identify the project manager in either consortium but if I have to think of one person neither is a scientist and they have a management mind set.

Q15. Can a one page Organigram be provided?

Have to be dug up.

Q16. What would you change with the advantage of hindsight?

Wait until we finish the mission. Today if I had to do it again I would go for one DPC not two. This Hipparcos model we adopted at the time dose not work for Planck. Today is too late for Planck.

Q17. What was the manpower/time estimate for your major software product?

Q18. What was the reality?

5 Software

Some of these of course could be answered with a yes or no but I am hoping for a little elaboration ☺

Q19. Are a set of software engineering standards used in the project (ISO,ECSS), is adherence checked ?

Both DPC claim to adhere to PSS05 Lite. The applied it loosely. They produced some documentation. It was a superficial effort though they were often written independently from the development with a life of their own not related to the development. Now we are making

an effort to get back on track – we have split the development one part is the launch critical part. The rest, the scientific part, it is accepted will be developed rather differently. In principle the launch critical part should adhere to something like ECSS.

Q20. Were standards mandated by a funding agency?

ESA mandated the DPCs adhere to some standards. They choose PSS05.

Q21. If standards were mandated state your opinion on their benefit to the project?

It would benefit the project to follow standards but we need to pare down to essentials as we are doing now with the launch critical software. The DPCs originally were too complex it was not possibly to know which parts needed to be tracked closely and which not. Standards need to be maintained at least for a part of the project.

Q22. Is a particular Software development methodology used in the project or parts of it (OMT, Booch, Waterfall)?

Nothing in particular. There is a certain approach, breadboarding , number of releases etc. but its not been adhered to.

Q23. If a methodology is used how was it selected?

Q24. Do you use a source code control system such as CVS?

CVS is being used successfully. Ask Adam for details.

Q25. Do you use a release management system such as ClearCase?

There is a release policy. In particular HFI have a clear policy.

Q26. Do you use a problem tracking system?

There is a system. This is used definitely in IDIS but less so in the DPCs.

Q27. Have you partnered with a major vendor for software production?

None.

Q28. Have you been able to use COTS (Common Off The Shelf) components in you system? (Even Freeware)? Did it save money?

The two DPCs use different databases. At the moment the Italians may use Oracle – at the moment they use flat files. They both tried Versant but it did not do the job. Which in some peoples opinions was not fair. The French have discarded that. Now they are trying to implement something on the Berkley DB system which is open source.

Process coordinator is developed in house at MPI. There is a basic layer which is common for the DPC which the process coordinator can deal with and each DPC has a more complex layer on top.

Q29. What is your main development language? (if you have one).

Varies from place to place – mainly C and C++, Java has not been so successful possibly due to the programmers not being so familiar with it.

Q30. How would you rate your processing in terms of difficulty, describe it a little?

Conceptually its not so complicated. There have been some changes in how people think about this. There was a feeling that we were hunting for an algorithmic break through to tackle the difficult bits (Map making and power spectrum extraction). That never came – the emphases now is to do as well as possible using approximation methods which are already know. So the emphasis is now not to do a perfect job but to do an approximate job which is essentially computer limited. So in a sense from that part, psychologically, the difficulty has gone, now it is just a question of doing as good as you can. Another change is that other parts which we thought were quite simple have turned out to me more difficult – dealing with time streams, systematic and instrumental effects. The perceived difficulty has moved from one part of the pipeline to another. For today's difficulties there is no breakthrough to be expected, it's a structural problem not only infrastructure not only to do with the quantity of data. The main difficulty is in the way you operate on the data keeping the throughput and keeping the people intervention where it should be. This complexity has to do with the understanding of the instrument.

6 Hardware

Q31. What kind of hardware system do you have, monolithic mainframe/supercomputer or cluster/distributed system or something else entirely?

LFI have a Beowulf system with 16 CPUs and they want to go to 30. HFI has got a parallel machine of some kind. The Italians have the idea to use the EGEE (Grid) – they are member of the consortium. I think you need some core you control – the grid can not be used operationally. They may work well for scientific processing. They have been using the NEARSC super computer machine in USA already quite a lot. For scientific processing there

will probably be quite heavy use of Grid or public computing resources but for the core there will be dedicated machines at each DPC.

Q32. Approximately how much processing power have you got?

Q33. How much disk space?

A lot !

Q34. How much disk and processing power was estimated for in the beginning of the project (if one was made)?

For the perfect job the estimates were so off scale and useless. So now we are computer limited and will do the best we can with the resources available.

Q35. Do you use a tape archive? If so is it still cost effective?

Q36. Have you partnered with a major hardware vendor? Was it successful? Did you ever feel locked in?

Q37. Anything else to add?

Management has been very tough. Michael's approach of letting ESA do data processing centrally or at least have some control is the correct approach. We made the decision a long time ago to totally let the PIs do the processing and ESA has no part of the processing and this was probably a mistake.

Appendix 7. Answers from the HEASARC facility.

1 Introduction

This is the questionnaire used to guide interviews about projects for the study. A summary of the answers is provided below under each question.

2 Background

Q1. What is your name and position in the project?

Tom McGlynn– Chief Archive Scientist for the High Energy Astrophysics Science Research Center.

Q2. May I record this conversation?

That is fine.

Q3. May I use your name in my final report?

Yes

Q4. Would you provide a brief summary of the facility or salient reference?

The HEASARC is NASAs main archive for high energy astronomy. The idea being that NASA provides a set of archives in wavelength domains and provide the long term storage for the data in those domains. So we have the HEASARC in the high energy domain IRSA in the infrared and MAST in the optical UV.

3 Costs

Q5. What is the budget of your facility?

$4 million a year

Q6. Is there a typical overrun or under spend on projects?

There is a degree of interplay with the missions, we share people and facilities so a there could be a small under or over run. Perhaps someone works for a few months for a project or vice versa.

Q7. How much was earmarked for Software development?

In context of total budget, it is all personnel these days. About half for software development and the other half for scientific support and documentation, running AOs etc. It's a bit fuzzy this number but roughly half is a good estimate.

Q8. Is there an over/under spend on software development, how much?

Q9. How much is earmarked for Hardware procurement?

HEASARC does not have enormous requirements – order of $100K per year.

Q10. Is there an over/under spend on Hardware, how much?

4 Management

Management is a tricky topic but one for great interest to me, especially for science projects. To put this in perspective again we are interested here in the management of the data processing and storage teams not perhaps the building of the entire instrument and general project.

Q11. What size has the team been over the lifetime of the facility?

15-20 person throughout its existence, roughly half scientist and half programmers, of course all of the scientists do some programming.

Q12. How many institutes are involved or is it all in house?

(Wil) You mentioned missions so those involve institutes ?

We have guest observer facilities and guest observer programs. We are actively archiving five missions and two under development, one which will launch tonight, we hope, (ASTRO2e). There is still development activity on some of the older missions. So almost ten altogether.

(WIL) Do you depend on those institutes?

The general interplay is that the data flows from the institutes into the HEASARC , they may or may not use the HEASARC facilities in their non mission critical activities but rarely use the HEASARC for mission critical activities. If the need data for that they keep at least the recent data online.

Q13. Has a particular management style been consciously adopted by the management?

No conscious management style.

Q14. Are managers formally trained?

There is some formal training – I have spent a week at WALOPS on training but that was for Goddard not for this facility as a whole. The contractors also have some training. But it is not a lot and not as much as is needed.

Q15. Can a one page Organigram be provided?

There is Nic White as head of HEASARC. There are no sub management. There are 2 govt scientists, 5 contractor scientist and 10 programmers not all fully funded by HEASARC.

Q16. What would you change with the advantage of hindsight?

I would change, the breakdown, having different contractors for science and programmers. I would have one contractor. Nic prefers separate people – this keeps NASA fully in control of everything. No contractor has control over another contractor.

Q17. What was the manpower/time estimate for one of your major software products? What was the reality?

Mostly level of effort. Mission software is different but not funded out of HEASARC. HEASARC inherited a system taken from ESTEC (Exosat browse) that was then improved on as needed. **(WIL gone back to Integral now)**

5 Software

Some of these of course could be answered with a yes or no but I am hoping for a little elaboration ☺

Q18. Are a set of software engineering standards used in the facility (ISO,ECSS), is adherence checked?

No formally approved standards – there are some standards in groups like for FTOOLS. There are agreed regression tests and distribution mechanism.

Q19. State your opinion on their benefit to the facility/projects?

Q20. Is a particular Software development methodology used in the project or parts of it (OMT, Booch, Waterfall, Unified, Iconix)?

No particular methodology. With 8 programmers and 5 million lines of code we are mainly maintaining and adding features as they are required by missions. We have no design documents, critical design etc.

Everything that is mission critical needs proper standards. If it works on the scientific viability of the mission it needs to be rigorous – we could have done with more rigor but not a lot.

Q21. If a methodology is used how was it selected?

Q22. Do you use a source code control system such as CVS?

For some of the software – CVS or nothing.

Q23. Do you use a release management system such as ClearCase?

No formal tool. Just procedural. 3.5 million lines of code are FTools which has a procedure.

Q24. Do you use a problem tracking system?

Couple. Bugzilla for some, FTools have their own home brew tool. Many bugs are fixed informally without tracking – outside FTools only perhaps 20-30% of bugs go through the formal system.

Q25. Have you partnered with a major vendor for software production?

no

Q26. Have you been able to use COTS (Common Off The Shelf) components in you system (particularly DBMS)? (Even Freeware)? Did it save money?

SYBASE, Bugzilla, lots of Linux free stuff, Nagios Monitors upness of the system. IDL. Old system had its own database and it was faster, much of the database code is for transactions – which we don't care about it means you need someone devoted to that then. This probably saves some money but it may not be a critical as you think – we have no gigarows, mostly 1million rows.

Q27. What is your main development language? (if you have one).

None in particular Perl, Tcl, C, Fortran, Java

Q28. How would you rate your processing in terms of difficulty, describe it a little?

Skyview does some – cut-outs and image registering. Some processing is done in ingest that tends to be driven by the data centres, we do checksums and such

6 Hardware

Q29. What kind of hardware system do you have, monolithic mainframe/supercomputer or cluster/distributed system or something else entirely?

Distributed cluster of Linux machines. All Red Hat so far. Still a few sun machines. Might be a few others. Users have mixture of Macs PCs

Q30. Approximately how much processing power have you got?

Take our standard 1ghz machine today, we have 4 dual processor 2ghz machines so 3 or 4 Gflops of power. Lots of other machines around – depends how far out you go out in the cluster We do have BEOWULF cluster for gravitational wave processing if we need it But not for the archive.

Q31. How much disk space?

Archive is about 6Tb and we have 20Tb of disk and about 10Tb more for user space. Bulk of this is on SANs.

Q32. How much disk and processing power was estimated for in the beginning of the project (if one was made)?

immaterial

Q33. Do you use a tape archive? If so is it still cost effective?

DLT for backups only. Started electronic distribution in 1990. Have not seen a tape for a long time.

Q34. Have you partnered with a major hardware vendor? Was it successful? Did you ever feel locked in?

no

Q35. Anything else to add on any topic?

The issue that is import is that the initial stand that you take on these topics of how you build things and such will last, it will make a permanent impression for good or ill. So it is very important whatever you do, its important to think though carefully what you want to do and stick with it. Take a place like ISDC which has a very rigorous approach, they really talk to each other, they have very clear and detail processes, they spend a lot of time doing this. You

finally get the thing after all those processes and it may not work. But the culture is established and it sticks, I don't think you can change it very easily later.