Managing Research Data in Big Science

Norman Gray, Tobia Carozzi and Graham Woan
SUPA School of Physics and Astronomy, University of Glasgow

2011 July 14, v1.1

Version	v1.1
URL	http://purl.org/nxg/projects/mrd-gw/report
Distribution	Public

This report was prepared as part of the RDMP strand of the JISC programme Managing Research Data.

Abstract

The project which led to this report was funded by JISC in 2010–2011 as part of its 'Managing Research Data' programme, to examine the way in which Big Science data is managed, and produce any recommendations which may be appropriate.

Big science data is different: it comes in large volumes, and it is shared and exploited in ways which may differ from other disciplines. This project has explored these differences using as a case-study Gravitational Wave data generated by the LSC, and has produced recommendations intended to be useful variously to JISC, the funding council (STFC) and the LSC community.

In Sect. 1 we define what we mean by 'big science', describe the overall data culture there, laying stress on how it necessarily or contingently differs from other disciplines.

In Sect. 2 we discuss the benefits of a formal data-preservation strategy, and the cases for open data and for well-preserved data that follow from that. This leads to our recommendations that, in essence, funders should adopt rather light-touch prescriptions regarding data preservation planning: normal data management practice, in the areas under study, corresponds to notably good practice in most other areas, so that the only change we suggest is to make this planning more formal, which makes it more easily auditable, and more amenable to constructive criticism.

In Sect. 3 we briefly discuss the LIGO data management plan, and pull together whatever information is available on the estimation of digital preservation costs.

The report is informed, throughout, by the OAIS reference model for an open archive. Some of the report's findings and conclusions were summarised in [1].

See the document history on page 38.

This report was prepared for, and funded by, the RDMP strand of the JISC programme Managing Research Data.

Release: 7b021525c2d7, 2012-07-17 09:55 +0100

Contents

0 Introduction

Astronomy is as old as human culture. Early agricultural civilisations required reliable predictions of the positions and motions of the Sun and Moon, in order to predict in turn seasons, tides, and river risings. Even in the absence of an extensive scientific model, these predictions relied on careful observations, preserved in the form of almanacs or ephemerides. Documents such as these associate astronomy with not only the first data archives but, since these artifacts still exist, also the oldest data archives in the world. Long-term digital preservation in astronomy is nothing new. We cannot resist saying more about this, in Sect. 1.8.

Astronomical archiving does however evolve, and in the last few decades both astronomy and particle physics have had to become leaders in large-scale data management.

Although astronomical images (now all born digital) have always been substantial in size, they have generally been reasonably manageable. Newer astronomical techniques – and we are thinking of 21st century radio astronomy and gravitational astronomy – are capable of generating truly challenging quantities of data; and particle physics has been generating, and addressing, intimidating data problems for decades. These problems cover both the management and preservation of large data volumes, as technical problems, and the preservation of the data's information content, on substantially varying timescales.

0.1 Project Background

The Managing Research Data/Gravitational Waves project (MRD-GW) is concerned with the data management arrangements of the LIGO Scientific Collaboration (LSC), and of the broader Gravitational Wave (GW) community. It is one of the six projects in the RDMP strand of the JISC Managing Research Data (MRD) programme [2].

The GW community was selected by the Science and Technology Facilities Council (STFC), at JISC's invitation, as a representative example of big-science data management practice – as we elaborate below, it has features of both the traditional astronomy and HEP communities, without being identifiable with either of them. While many of the specifics, below, relate to this community, we believe much of the discussion is relevant to the other disciplines. Here, we are focusing on the big-science projects which receive strategic support from STFC, rather than the smaller-scale projects funded by specific research grants, since it is these large-scale projects that are distinctive about STFC-funded research. We assume that the outputs of the smaller projects will be managed through disciplinary repositories, in a manner which more closely resembles that of other research councils.

The MRD-GW project exists to inform three sets of stakeholders:

- Although the Joint Information Systems Committee (JISC) and the Digital Curation Centre (DCC) have extensive experience with digital libraries and digital curation in general, there are problems specific to ‘big science’ data which JISC would like to better understand.

- The Research Councils have recently started to require bidders to include a ‘data management plan’ within project proposals. However there is no consensus on what such a plan should look like for science funded by the STFC. The US National Science Foundation (NSF) has recently placed binding requirements on projects to produce data management plans [3].

- The LSC community has considerable internal software and administration experience, and has solved a large number of data management problems

focused on large-scale data storage and transport. However there is an awareness that (partly because there have been no immediate imperatives to do so) there was until recently no published plan for a long-term data archive.

The existence of these three groups is reflected in the overall structure of the document.

This project's context also includes the broad Virtual Observatory (VO) movement, which aims to develop standards and areas of consensus which help scientists have ready access to astronomical data across sub-disciplines and wavelengths. All the stakeholder groups have interests in the success of the VO movement.

The project aims to bring together two sets of practice, namely the long-term digital preservation perspectives represented by the OAIS reference model in the abstract and the DCC in particular, and the very considerable experience of practical large-scale data management, embedded within the LSC community.

For OAIS, see [4] and Sect. 3.1.1; in this report the 'DCC' is the JISC Digital Curation Centre, not the LIGO Document Control Center.

0.2 How to read this document

This document is organised into three main sections, broadly corresponding to the three audiences we are addressing.

Sect. 1 is about data management in big science. It is addressed to the JISC and to the data preservation community in general, and is intended to illuminate the ways in which scientists in these areas have distinctive data management requirements, and a distinctive data culture, which contrasts informatively with other disciplines.

Sect. 2 is primarily addressed to STFC and other similar funders of this type of science. It is concerned with the responsibilities which are imposed on funders by the wider society, and which are passed on to the funded through requirements on the governance of projects and the availability of data. The recommendations here are concerned with how best to express these responsibilities.

Finally, Sect. 3 is primarily addressed to the LSC, as a proxy for similar big-science projects. The explicit recommendations here are intended to be of as much interest to projects, as actions they may wish to take, as to funders, as behaviour it may be prudent or productive to require.

0.3 Working with communities – pragmatics

This report is the result of a fruitful collaboration with the GW community. It may be useful to note some of the features of the project, and the community, which contributed to this.

- The project team, as part of Glasgow University, has current involvement in the community, and the project director (Woan) is a senior figure there.
- The LIGO community is already aware of the general need for data management, and the specific need for preservation (see [5]).
- The project personnel have relevant scientific background, and are to some extent in the position of being informatics-for-astronomy specialists (ie we're 'insiders').
- The community is large and (via studies such as [6]) has some experience of being 'studied'.
- The existing LVC workshop series meant that we could contact relevant people easily in a context where newcomers were expected, and we didn't have to add our own data management workshop.

1 Data management in Big Science

1.1 LIGO in perspective: LIGO, big science, and astronomy

What is 'big science'?

Big science projects tend to share many features which distinguish them from the way that experimental science has worked in the past. Such projects share (non-independent) features such as:

big discoveries These projects are expected to be amongst the most important ones of their generation. Although there is very high confidence that their headline science goals (for example the Higgs and GW searches) will be successful, they are also expected to produce long lists of unexpected results, and a broad range of engineering spinoffs.

big money These are decades-long projects, supported by country-scale funders and billion-Euro budgets (the total budget for the LHC is around three billion Euros, not including the detectors, nor the personnel and hardware costs directly supported by country funders, which cost between one and two times that sum).

http://askanexpert.web.cern.ch/AskAnExpert/en/Accelerators/LHCgeneral-en.html

big author lists The projects involve collaborations of hundreds of people (the LSC author list runs at around 600 people (cf Sect. 1.6.1), and the LHC's ATLAS detector author list is around 3000).

big data Enhanced- and Advanced-LIGO (for example) will produce of order 1 PB yr^{-1}, comparable to the ATLAS detector's 10 PB yr^{-1}; the eventual SKA data volumes will dwarf these.

big admin MOUs, councils, workshop series.

big careers Individuals may make the journey from PhD to chair on a single project.

There is a discussion of the features of 'big science', and LIGO's progress towards that style of working, in [7], with an extended history of the sub-discipline in [6].

Because of the large costs involved and because there is usually little immediate commercial value in this research (though of course there are substantial long-term economic payoffs for the investing countries), these large projects are funded at the national or international level, so that taxpayers are the ultimate stakeholders. Even putting aside the scientific and scholarly need for adequate data preservation, these national investments make it necessary for funders both to demonstrate that projects are being efficiently exploited to produce macro-economic value, and to make the data products available for public use. We discuss open data in Sect. 2.2

1.2 Data volumes

In the context of larger-scale projects, a 'science run' is a period when the equipment is run more-or-less continuously, gathering scientifically useful data. Between science runs, the experiment will either be down for maintenance, or on a planned 'engineering run'; data from engineering runs is generally stored, but is not expected to be useful to scientists.

ATLAS, one of the two larger LHC detectors, stores 3 PB yr^{-1} by itself; see http://atlas.ch/pdf/atlas_factsheet_4.pdf for some entertaining numbers.

The most immediate problem with data curation and sharing in these scientific areas – though in the end not the most significant one – is the data volumes involved. The current volume of LIGO data is of the order of hundreds of terabytes, and the data rates is expected to grow, over the course of the project, from its current 100 TB yr^{-1} to around 1 PB yr^{-1} (see Table 1 on the facing page, which shows the variation in data size for science runs three to six).

LIGO is just one of several other existing or planned big physics projects, including the LHC, the Square Kilometre Array (SKA), and various European Space Agency (ESA)/NASA space missions. In comparison with these projects, LIGO's data handling requirements are relatively modest. The LHC will have data volumes of tens of PB yr^{-1} Further in the future, the SKA (which is due to be com-

	S3	S4	S5	S6
L0	57	32	816	261
L1	8.24	4.04	119	76
L2	1.55			
L3	0.97	0.86	9.70	3
(duration/day)	70	29	695	482

Table 1: LIGO data set size estimates in TB, and run lengths in days, for science runs three to six ('Sn'), and various data types (size data taken from [8]; there were a total of six science runs in LIGO; L0 is the run's raw dataset, and L1 to L3 are progressively reduced).

missioned around 2020) has predicted requirements up to 1 Tbit s^{-1} locally and 100 Gbit s^{-1} intercontinentally; this involves transporting, though not necessarily storing, around 1 TB min^{-1} or 0.5 EB yr^{-1} [9]. This is 0.05% of the predicted 1 ZB yr^{-1} *total* worldwide IP traffic for 2015 [10].

1 PB is 1000 TB; 1 EB is 1000 PB; 1 ZB is 1000 EB; note that the unit B refers to bytes, not bits

Large-scale physical science experiments have long produced significant data volumes, but in recent years datasets appear to be increasing in volume and in complexity at an overwhelming rate, and this may present a qualitatively different data management problem. This is sometimes described in rather apocalyptic terms – as a 'data deluge' or the like – and some of the challenges and opportunities are described in [11].

1.3 Data management styles in the physical sciences

It seems useful to discuss, here, some of the distinctive features of data collection and management in the experimental physical sciences, since these have an impact on both the expectations for, and the problems with, the data.

Big-science research projects have a number of relevant common features:

Large data sets Such projects' data sets are 'large' in the objective sense that the projects are typically so greedy for data storage, that their holdings are near the edge of what it is technically feasible to store and transport.

Innovative data management As part of the response to their need for large data volumes, big-science projects are often extremely innovative in their solutions to data management problems, to the extent that they are willing to work with experimental filesystem types, or adapt and extend operating system software or network transport protocols (see `http://lcg.web.cern.ch/` to get an impression of the scope of development efforts here).

Specialised software Because the instruments and their data sets are so complicated, these projects typically generate large custom data analysis software suites. These may require specialised and unwritten knowledge to use, and therefore appear to represent a significant software preservation challenge.

Beyond the substantial software engineering challenges described above, the physical sciences tend to have few 'IT' problems, since the communities contain plenty of people with sufficient technological nous to address essentially all day-to-day computing-related problems, and these communities are therefore generally reasonably well-organised with regard to backups, storage, and basic file sharing (see also the discussion of technological readiness in Sect. 3.5). At the same time, however, the communities are rather conservative from the point of view of a computer scientist, and sometimes rather informal from the point of view of a software engineer. That is, the attitude to custom computing solutions is very similar to the attitude to custom lab hardware: it may need to be creative

and experimental, but never for its own sake; it must be stable, but is never frozen; it is accurately made, but rarely polished. The analogy with lab hardware and software holds to the extent that, in the LHC community, data management groups are regarded as detector subsystem groups; that is, they have the same general status as the magnet or accelerator engineers, and expected to produce agile and innovative computing services very different from the more routine, lab-wide, provision of CERN IT services.

One of the DCC researchers, commenting on this report, quoted a GridPP survey respondent commenting that the LHC computing task: "Providing resilient services that maintain access to data for the experiment users 24/7 – services are complex, bleeding edge, and are constantly being updated. Controlling that process, whilst also maintaining service up-time is very challenging".

The result of this is that lab software represents functional solutions to immediate-term problems, generally with flexibility enough to respond to medium-term problems, but without much attention being given to the imponderables of the long term, after the experiment has completed. It is precisely these Long Term preservation questions, in the OAIS sense of more than one technology generation, that are the concern of this report.

There is plenty of prior art in this area. See reference [12] for a review of data management practice in a variety of scholarly areas, which additionally covers several proposed life-cycle models, and analysis techniques. There is a similar overview in the PARSE.Insight case-studies report [13], which examines data management practice in HEP, earth observation, and social science and humanities. These case-studies were conducted via interviews, and participation in ongoing efforts within the communities. The same project produced a gap analysis and roadmap, which make valuable reading.

This is a good place to stress that 'big science' generally handles its data well, and can even be regarded as exemplary (compare Sect. 3.5). There are a few features which naturally encourage good data management practice in the large-scale physical sciences.

- These are often relatively well-resourced projects, with plenty of computing experience and lots of engineering management. There is lots of obvious infrastructure in the development of a large collaborative experiment, which gives data management an obvious budgetary home, where it is not competing with funding which directly supports researchers.
- Astronomy and HEP projects have always produced 'large' data volumes: this makes *ad hoc* data management manifestly unattractive, and encourages explicit data management planning and discipline.
- The scale of these experiments means that they tend to be shared facilities providing documented services to their users, so that documented interfaces and SLAs are natural.
- These projects rarely if ever produce commercially sensitive data, so that the confidentiality concerns are well circumscribed, concerning professional priority rather than IPR or other financial worries.

Although these features are to a greater or lesser extent specific to this type of science, they have given rise to the notions of *data products* and explicit *proprietary periods*, which we believe would be useful in other areas, and which we discuss in Sect. 1.11.

Although it is GW data which is our nominal focus in this report, it is convenient to first describe general astronomy data, then distinguish that from High Energy Physics (HEP) data, which has a somewhat different data culture, and then describe how the GW community, which is in many ways intermediate between the two, handles its data.

1.4 Astronomy data

Astronomy (excluding GW astronomy for the moment) has probably the most straightforward data management practices in the physical sciences. When an

optical telescope takes an image (or a spectrum, which for our present purposes is technically equivalent to an image), either as part of a systematic survey of the whole sky or as a pointed observation requested by an astronomer, the image is typically moved from the telescope's detector straight into its archive, from where it can be later retrieved by the astronomer, accompanied by automatic or manually-added metadata.

Non-optical astronomers (covering the rest of the spectrum, from radio to gamma rays), and most satellite missions, have a somewhat more complicated route from observation to image, and a broader set of data products, but have essentially the same model, and the same discipline and expectations around archives. From the point of view of data management therefore, we can elide the differences between the various branches of astronomy. Gravitational wave and neutrino astronomy, in contrast, are not studying the electromagnetic spectrum, and partly as a consequence their study more closely resembles particle physics (see Sects. 1.5 and 1.6 below).

Most large telescopes, satellites and instruments operate partly or exclusively according to a model in which astronomers are awarded 'telescope time', ranging from a few hours to a few nights, as the results of competitive bids closely analogous to grant bids. The resulting data generally has a proprietary period, extending for perhaps 12, 18 or 24 months after the data is taken, during which only the observer who requested it can retrieve it, but after which it automatically becomes retrievable by anyone ('embargo' would be a better term, though unconventional). Similarly, instruments built by consortia generally have proprietary periods during which the data is only available to consortium members. The proprietary periods are partly for the benefit of the consortium individuals – it is their reward for the initiative and possibly decadal effort of building the instrument – but they are also a pragmatic reflection of the length of time it may take to calibrate and validate acquired data, ready for deposit in an open archive. As a result, the lengths and terms of proprietary periods are the subject of negotiations between the instrument builders and their ultimate funders, though the negotiations are always about the length of the delay before a general data release, and never question the necessity for the release itself.

An 'instrument' in this context is the light-sensitive detector attached to the telescope or satellite optics (the camera, in effect). It is replaceable or swappable, and regarded as a separate piece of engineering from the telescope. The days when observers would travel to the telescope carrying their own instrument are now largely past.

NASA missions now typically have 12-month proprietary periods, but this has varied historically, and for example the 1990 COBE mission, which included significant technological novelty, and whose performance was therefore rather unpredictable, had a 36-month proprietary period.

Not all instruments have formal release plans, and the proprietary periods that exist may be adjusted informally. Caltech is one of the few private institutions which is rich enough to own, or have a significant share in, world-class telescopes (Palomar and Keck). It has no declared policy on data management or data sharing, beyond a broad tacit expectation that data will be published as appropriate for normal scientific practice. As a second example, during the 'science demonstration phase' of the commissioning of the Herschel telescope (that is, the last commissioning phase, verifying that the science goals were achievable), the instrument team invited observers to nominate part of their scheduled observations to be performed early, during this still-experimental commissioning phase. When the observations proved successful (as they generally were), the observing teams were given the choice either of making the data immediately public, in time for the opening of the Herschel archive and a journal special issue, and having the observation time re-credited to them; or else retaining the 12 month proprietary period without the re-credit.

See 'Herschel Observers' Manual' §1.1.4, `http://herschel.esac.esa.int/Docs/Herschel/html/ch01.html`; thanks to Haley Gomez for bringing this to our attention.

Image data is the archetypal astronomy data, and is generally stored as files, but another important category is the astronomical 'catalogue', of object positions, spectra and other properties, usually stored in relational databases. Astronomical archives range from quite small ones (at one extreme, a small specialised instrument may have its 'archive' consisting of a file server looked after by a grad-

uate student) to very large professionally managed archives which are both the primary sources of some data sets, and mirrors of others.

Astronomy data is potentially very long-lived. Although astronomers are naturally drawn to the newest instruments with the greatest sensitivity, it is not unusual to draw on relatively old archive data. In most cases, this will be still be born-digital data, but digitised versions of century-old astronomical plates are used in precise astrometry, and to identify the precursors of supernovae and other one-off events (see for example the Edinburgh SSA, further discussed, with background, at [14]; and a more discursive account of plate scanning, including discussion of some of the archival challenges, in [15, 16]). Even babylonian and ancient chinese astronomical data has been used for contemporary science, helping measure the rate at which the earth's spin rate, and thus the length of day, is changing [17]; similarly, 3 000-year-old egyptian data has been used to measure the change in the orbital behaviour of the three stars in the Algol system [18]. The cosmos changes slowly on our timescales, so that the great majority of astronomical observations are repeatable; the exceptions are those cases where long time-bases are necessary (precise astrometry) or where the object of study is a one-off, and therefore unrepeatable, event such as a recent or historical supernova.

http://surveys.roe.ac.uk/ssa/

Astronomy data is also intelligible in the long term: although untranscribed babylonian tablets can only be read by specialists, contemporary astronomers can basically understand the data published in Kepler's 1627 *Rudolphine Tables*, and with some assistance can understand the content of the 11th- to 12th-century Toledan Tables [19]. Although biologists might be able to make similar claims with respect to, for example, Linnæus's observations, it is hard to find equally long-lived data in the physical sciences, or born-analogue physics data where there is a similar contemporary pressure for digitisation.

There is essentially no file-format problem in (electromagnetic) astronomy, since the Flexible Image Transport System (FITS) format is universal [20, 21]. Though not perfect, this is a relatively simple and well-defined format, combining binary or table data with keyword-value metadata.

Astronomical data also has a well-developed notion of *data products*. These are datasets which contain, not raw data, but data which has been processed to a greater or lesser extent. We can distinguish at least three levels of data in this context; most large instruments will have more than one level of derived products.

Raw data This is the lowest-level data, consisting of the direct output of a detector or other instrument, or the raw satellite telemetry. This data is made meaningful only by processing with software which is to some extent specific to the equipment in question. Though it will be preserved as a matter of course, it is rarely published, nor used by, nor useful to, other researchers, except in unusual circumstances. In the case of a particularly subtle effect – or less commonly, a debate over a theoretical analysis or calibration – a researcher might return to the raw data, but this will generally be done with the collaboration of the instrument scientists, and may be otherwise infeasible, to the extent that any results obtained without such insider knowledge might not be believable by the broader community.

Data products After it is gathered, (raw) data must be processed ('reduced') to turn it into scientifically meaningful numbers (interpreting engineering or telemetry data streams, and calibration) and to remove various instrumental and observational artefacts. Data products are usually made available in standard formats (in astronomy, generally FITS files), whereas raw data, if it is made available at all, may well be in an instrument-specific form.

Publications Sitting above the data products is a class of high-level outputs, in-

cluding scientific papers, and other peer-reviewed outputs such as published catalogues. Journal articles are curated at publisher sites and the Astrophysical Data Service (ADS), and article preprints at the arXiv (cf Sect. 1.9). Modest volumes of data can be published as digital appendices to journal articles in, for example, Astronomy and Astrophysics Supplements; these are curated at the journal and at VizieR.

It is the data products which are the outputs which are sufficiently free from observational artefacts to be the starting point for scientific analysis (high-level products are sometimes referred to, informally, as 'science data'), and which represent the class of data which is naturally archived, most carefully documented, and which will eventually be made public. There may be multiple levels of data products, with lower-level products carrying more information, but using which requires more detailed knowledge of the subtleties of the instrument and its processing pipeline. To a much greater extent than is true for HEP data, for example, the highest level astronomy data products are both useful and generally intelligible – everyone is, after all, looking at the same sky – but researchers will often use intermediate-level products, if they can invest the time to learn about them, or have collaborators who have experience with them. Those researchers who are more intimately involved with an instrument will be comfortable using lower-level data products, because they will have the knowledge which enables them to run, or experimentally re-run, the pipelines in a scientifically meaningful way. That said, in OAIS terms, astronomical data can be characterised as having a broad Designated Community and well understood Representation Information.

Publications are in the province of libraries and similar repositories, and are not considered further in this report.

Optical astronomy (that is, with observations made using visible light) has the most straightforward data, so that the distinction between raw data and data products is slight to the point of being rather artificial: astronomers reusing optical data would expect to recalibrate the raw or nearly raw data, and would not anticipate having difficulty doing so.

We conclude with some examples: A typical telescope archive is the UKIRT archive at `http://archive.ast.cam.ac.uk/ukirt_arch/`; there are several image and spectrum archives at the Royal Observatory Edinburgh's Wide Field Archive Unit; and there is a large collection of catalogues available at Strasbourg Data Centre (CDS) (see Sect. 1.4.1 below).

`http://www.roe.ac.uk/ifa/wfau/`

The ESA Hipparcos astrometry mission flew between 1989 and 1993, and produced a high-precision catalogue of 100 000 stars [22]. The catalogue is available online as queriable databases at ESA and CDS, as CDs, and as PDFs which match the catalogue's 17-volume printed version. The printed version is an interesting case: as discussed in the catalogue (vol. 1, §2.11.3), the printed pages are designed with a per-page checksum, to help with re-scanning the catalogue from paper, in the presumed-likely future case that the digital version becomes unreadable and only copies of the paper book survive. The Tycho catalogue, from the same mission, comprises around 20 times the number of stars, at lower precision, and is only available online.

`http://www.esa.int/science/hipparcos`

`http://www.rssd.esa.int/index.php?project=HIPPARCOS`

There is some discussion of preservation costs in Sect. 3.4.

1.4.1 Strasbourg Data Centre (CDS) as a disciplinary repository

CDS is a large disciplinary repository for astronomy [23]. It stores a broad range of catalogues, of various sizes, in its VizieR service (see [24] and `http://vizier.u-strasbg.fr/`) and provides a large librarian-curated collection of data from, measurements of, and references to, individual astronomical objects. It cooperates closely with ADS.

CDS was created, and is supported, by the french agency in charge of ground-based astronomy – first CNRS/INAG then CNRS/INSU – as a joint venture with

Strasbourg University. The main support is through permanent positions from the CNRS/INSU and the University (researchers, computer engineers, and specialised librarians), with additional contracts supported by funding from various sources.

CDS is administratively located within a research structure, Strasbourg Observatory, providing an active research environment for CDS astronomers. The preservation aspects have never been separated from the provision of services and the maintenance of local expertise on data management and preservation.

We are most grateful to Françoise Genova, of CDS, for this discussion of CDS's history and support.

This can be seen as an example of a very successful disciplinary repository. There appear to be several key features of this success.

- CDS has established, and actively maintains, international leadership in the curation of astronomical data, by virtue of collaborating widely and investing effort in projects (such as the International Virtual Observatory Alliance (IVOA)) which support and promote data sharing.
- As a result of the intimate relationship between the repository, the observatory and the university (to the extent that the boundaries between the three can seem rather vague to outsiders), CDS personnel have practical knowledge of how their data is used, and what researchers need.
- The core funding for CDS comes from the french state, but it is conceived as an internationally visible project.

1.4.2 Collaborations in astronomy

The most visible collaborations in astronomy are large terrestrial and satellite-borne telescopes and other instruments. At the risk of oversimplifying, these are generally not the many-person collaborations usual in GW or HEP physics, but are instead facilities created by space agencies or consortia of national funders. Although they are highly innovative leading-edge facilities, they are not seen as *experiments* in the same way as LIGO or the LHC are (massive) items of specialised hardware built to answer a delimited set of scientific questions. They are instead *observatories*: the data management in these facilities is part of their general operating infrastructure, and the research and research data they produce is 'owned' (at least in an academic rather than a legal sense) by the scientist *users* of the facilities, rather than the facility itself.

Astronomy does however have a variety of data-analysis collaborations. These are semi-formal collaborations concerned, mostly, with multi-wavelength studies of multiple archives, and include for example UKIDSS-UDS, the Herschel Atlas collaboration, HerMES and GAMA. These have between 20 and 60 collaborating members scattered over perhaps a dozen institutions but, crucially, no 'corporate' existence, and little or no direct funding. Instead, they are funded indirectly via individual fellowships or rolling grants: participation in the collaboration might be a strong feature of a grant application, but it is not the collaboration as such that receives the direct support. They do have governance structures, but these tend not to be particularly formal, because they remain small enough that there is little perceived need. These collaborations exist to derive high-value derived data products from the lower-level data products of the archives they are analysing (for example Herschel is an ESA observatory mission: this means that individuals can bid for observations, but that ESA does not have it as part of its remit to provide more than minimally reduced science data).

http://www.nottingham.ac.uk/astronomy/UDS/, http://www.h-atlas.org/, http://hermes.sussex.ac.uk/, http://www.gama-survey.org/

The collaborations distribute their results in papers, and associated datasets; they typically build archives to support and distribute their work, but there's no expectation (beyond the usual cooperative academic norms) that they will help others work on the data, or release it. It is hard to see how there could be such an expectation, much less an obligation, since they receive little direct funding, and

their indirect funding comes from a multinational set of entities with potentially very different Data Management and Preservation (DMP) policies.

1.5 High Energy Physics data

Astronomy is essentially an observational science: telescopes, their optics, and the detectors which hang off them, are constructed to create a path from nature to data which is as nearly as possible unmediated. This means that it is both reasonably obvious what things are to be archived, and that the nature and processing of observational artefacts are well and commonly understood. This means that astronomy, unusual in the physical sciences for needing to preserve data long-term, is in the happy position of having its data readily preservable.

HEP data is different. HEP is a participative science, where objects ranging in size from electrons all the way up to nuclei are disassembled, and data about the messy results of this disassembly is examined to retrieve information about the interior structure of the original. This reconstruction from collision data depends on a shifting engineering understanding of rivers of data, out of instruments which are one-off works of art, designed and assembled by a thousand-strong community, close-packed into a detector the size of a small cathedral, attached to a machine with its own postcode.

The result of this is that HEP data analysis is rather tricky, with many steps between data and science, each of which depends on software which encodes a detailed understanding of the data's provenance. In consequence, although HEP data is typically distributed with multiple levels of reduction, almost none of these levels (with the exception of formal publications) are straightforwardly suitable for long-term preservation. This is because interpretation of this data is heavily dependent on software, the use of which requires detailed experimental knowledge which it may be infeasible to preserve. In OAIS terms, the designated community is tiny because the Representation Information is hugely complex.

In addition to this, HEP data has a considerably shorter shelf-life than astronomy data, as discussed above. In contrast, old HEP data is typically made redundant by new data, obtained from more powerful accelerators. Also in contrast to astronomy data, HEP data is not expected to be generally intelligible for very long: two- or three-decade old data might potentially be useful or intelligible, but much beyond that would count as archaeology. At the risk of being whimsical, we can compare the roughly millennial lifespan of astronomical data with the roughly three-decade lifespan of HEP data, and conclude that the latter goes 'off' about 30 times faster than the former. Although facilities make very considerable efforts to manage data safely while an experiment is running, there is little real pressure to preserve HEP data into the long term.

Of course, things are not quite as straightforward as that in fact. (i) The LHC gains interaction energy at the expense of a messier collision, so there are potentially some features that will be detectable in one dataset (for example the HERA p-e data) which would not be findable in the LHC. While interaction energy is the most prominent metric of an accelerator's performance, it is not the only one, so that larger accelerators will not render smaller ones obsolete as inevitably as we may have suggested above. Similarly to this, (ii) data reduction errors may be dominated by theoretical uncertainties rather than experimental ones, and these will only be improved, and the data re-reduced, after the experiment is over. Finally (iii) there are no accelerators bigger than the LHC currently scheduled, so that this dataset may remain the highest-energy one for a relatively long time. The archaeology is illustrated in [25] and the problem further explored in [26], which also discusses the HEP community's developing plans for data preservation. Qualifications notwithstanding, the overall timescales in HEP are shorter than in astronomy, and the solutions described in [26] are concerned with prolonging a continuous low-level relationship with a dataset rather than being able

to return to a dataset cold.

Unlike astronomy, HEP has for the last few decades been organised into larger and larger collaborations, and these collaborations have developed intricate, and socially fascinating, cultures for managing this. The two larger instruments at the LHC, ATLAS and Compact Muon Solenoid (CMS), each have author lists of order 3 000 people, so that the various CERN collaborations account for around 10 000 research-active individuals. There is extensive discussion of the history and structure of the LHC collaborations in [27] and in the outputs of the PEGASUS project, but many of the collaborations' relevant organisational features are echoed in the GW community: this is discussed in Sect. 1.6.1 and we do not discuss them here.

http://www.pegasus.lse.ac.uk/research.htm and in particular [28]

1.6 Gravitational wave physics

The gravitational wave community has astronomical goals, but in the scale of the LIGO project, and in the amount of novel technology involved, as well as in the fact that many of the personnel involved came originally from a HEP background, the project's culture more closely resembles that of a HEP experiment than of an astronomical telescope. We discuss some specific features of LIGO data in [29]; here we discuss where GW data, and the discipline's organisation structure, fits on the spectrum between astronomical and HEP data.

1.6.1 Gravitational wave consortia

There are three principal sources of recent GW data available to UK researchers: LIGO, GEO600 and Virgo. There are other detectors which are either smaller efforts (in terms of consortium sizes), which have stopped taking data (TAMA-300), or which are still at the planning stage. See [30] for an overview of current detectors, and of detector physics.

LIGO Lab is a collaboration between Caltech and MIT, which designs and runs three interferometers in Hanford, WA, and Livingston, LA, in the US. GEO is a German/British collaboration, which runs the GEO600 interferometer. The three LIGO interferometers were shut down in October 2010 to refit for Advanced LIGO (aLIGO); the GEO600 interferometer is still currently running. The LSC is the result of a network of Memoranda of Understanding between LIGO Lab (or more loosely the LSC) and multiple other institutions of various size. These relationships involve hardware, resources, and data access of various types. Most typically, the resources in question are personnel, and an institution such as a university physics department, which wishes access to LIGO data, will contribute in return fractions of staff from permanent staff, through post-docs, to PhD students, for a broad spectrum of activities including data analysis, instrument fabrication and shift-work in the detector control room. However in some cases, the MOUs are concerned with data swaps, and set up limited data releases with other scientists: for example, there are a few MOUs between the LSC, Virgo and other observatories, which describe what data is to be shared, in what volumes, and the outline authorship arrangements for any subsequent papers. GEO's MOU describes a particularly close relationship with LIGO Lab, but most of the MOUs are broadly similar to each other, and the process of creating one is by now streamlined. In total (as of June 2010), the LSC consists of a little over 1300 'members'; of these, 615 spend more than 50% of their time dedicated to the project and so have a place on the LSC author list.

The MOU which created the LVC is at [31], but MOUs are not routinely made public.

The definition of LSC membership is included in [32] and the construction of the author list in [33].

The term 'LVC' is not an initialism. It colloquially refers to the data-sharing agreement [31] and joint meetings between the LSC and the Virgo Collaboration. Though there are 'LSC/Virgo collaboration groups', there is no formal big-C Collaboration.

The term 'LIGO' has a number of not quite equivalent meanings: sometimes it refers to LIGO Lab, sometimes to LIGO Lab plus the LSC, and the phrase 'LIGO detectors' is generally understood to refer to the LIGO Lab and GEO detectors.

The Italian/French Virgo consortium has its own detector and analysis pipeline, and has a data-sharing agreement with the LSC, represented by the LVC. As with

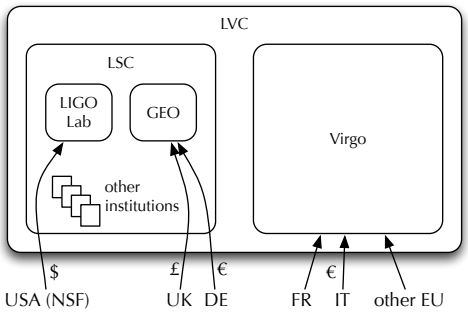

Figure 1: The relationships between various GW consortia.

LIGO, the Virgo detector will shut down between 2011 and roughly 2015. Virgo has 246 members (with a slightly different definition from the LSC), and GEO600 around 100.

There is an attempt to summarise these relationships in Fig. 1.

These experiments have a common purpose: they exist to detect signatures of gravitational waves, which are confidently predicted by the General Theory of Relativity, but the actual observation of which would be a major scientific event (there exists an LSC data processing flowchart which includes the not entirely serious branch "Call Stockholm!").

Gravitational waves are sufficiently weak, however, that the existing equipment will not become sensitive enough to have a good chance of detecting them until after its refit, which began in late 2010 (when the project entered the phase known as aLIGO), and which is scheduled to be completed when the new detectors are commissioned in 2015.

1.6.2 GW data

Although the consortia have (as expected) announced no detection so far, they nonetheless produce a large volume of auxiliary data, representing background and calibration signals of various types, and this, together with the core data, means that the LSC collectively produces data at a rate of approximately one $\mathrm{PB\,yr^{-1}}$.

We can readily identify the levels of data which were discussed in Sect. 1.4:

Raw data The lowest-level GW data consists of the signals from the core detectors. This data is made meaningful only by processing with software which is completely specific to the detectors in question. This is stored in 'frame format', which is a very simple format intelligible to all the primary data analysis software in the community, and which is multiply replicated across North America, Europe and Australia. Although the disk format is common, the semantic content of the raw data is specific to detectors and software, so that preserving it long-term would represent a significant curation challenge.

Data products The raw data is processed into calibrated 'strain data', which is the data channel in which a GW signal will eventually be found (this is possibly, but not necessarily, also held in frame format). This is the class of data products which will eventually be made public. Unusually, it turns out that GW raw data is in a semi-standard format, and the data products are specific to the analysis pipeline which produced them.

Publications Sitting above the data products is a class of high-level data products, scientific papers, and other peer-reviewed outputs. The GW projects have announced no detections of gravitational waves, but have nonetheless

produced a broad range of astrophysically significant negative results [30, §6.2].

As with the general astronomy data products discussed in Sect. 1.4, the distinction between the 'raw data' and the 'data products' is that the latter datasets, alongside their supporting documentation, will be available for use and reuse by scientists who do not have an intimate connection with, and knowledge of, the instrument.

Both the 'data product' and 'publication' groups are broad classes of objects. The practical boundary between them is clear, however: what we are calling 'publications' are entities such as journal articles or derived catalogues whose long-term curation is not the responsibility of the LSC data archive, though they may be held in some separate LSC paper archive, which is as such out of scope for this project.

1.6.3 Gravitational wave data releases

Because the LSC has not announced the detection of any signal so far, and because the data will remain proprietary to the consortium until well after such an announcement, there are no distributed data products so far, and so the issues surrounding formats and documentation have not yet been addressed. However it is the eventual public data products which are the highest-value outputs from the experiment, and which are the products which it will be most important to archive indefinitely.

At present, LIGO data is available only to members of the LSC. This is an open collaboration, and research groups which join the LSC have access to all of the LIGO data. In return, they contribute personnel to the project (including for example people to do shift-work manning the detectors), and accept the collaboration's publication policies, which require that all publications based on LIGO data are reviewed by the entire collaboration, and carry the complete 800-person author list. At present, and in the future, data which is referred to by an LSC publication is made publicly available. See Sect. 3.3.2 for further details on LIGO's DMP plan.

http://www.ligo.org/about/join.php

1.6.4 Summary: big-science preservation challenges

In the three sections above, we have tried to describe both differences and commonalities between three large-scale scientific disciplines. Possibly the biggest difference between the three areas is that high-level astronomical data products are much more generally intelligible than even the highest-level HEP products. In each case, however, we have a ladder of reasonably well-defined data products, with each rung generated from the lower ones by sophisticated data reduction pipelines.

The situation is not as rosy, from the point of view of long-term preservation, as this account may suggest. Because the pipelines have developed organically over a number of years, under the influence of experience with earlier versions and increased understanding of the instrument, the knowledge they represent is sometimes encoded within them in a less structured way than would be desirable. Sometimes, metadata is encoded in filenames, or in configuration files, or wikis, or even private emails. Of course, one could simply argue that this information should be documented better, but it would be hard to argue that the costs of this work would be justifiable, to service a future theoretical need that few believe would even become an actual one. In consequence, although the resulting data product will be regarded as perfectly reliable, it may be infeasible to redo the analysis other than by preserving and rerunning the pipeline software (even if it were feasible, it would be prohibitively expensive, and rarely seen as valuable;

see also Sect. 2.4). For this reason, software preservation has some role in the overall data preservation strategy. However it is not clear to us what this role should be, and the thorny issue of software preservation is addressed at greater length in Sect. 3.2.

1.7 A contrast: social science data

It is possibly instructive to contrast the data management practices discussed here, with the very different problems faced by data managers in the social sciences. In [34], the authors survey a number of social science projects, with a particular focus on two large (for the social sciences) programmes funded by the Economic and Social Research Council (ESRC) (the UK social science research council) with substantial responsibilities for data preservation and sharing.

This work was part of the 'Data Management Planning for ESRC Research Data-Rich Investments' project (DMP-ESRC) (http://www.data-archive.ac.uk/create-manage/projects/JISC-DMP), funded by JISC, like the present project, as part of the Managing Research Data programme.

For the ESRC projects, the artefacts being stored are simple things, at the level of Content Information: they are conventional Word documents and audio files, rather than the heavily structured and still somewhat experimental big science data objects. The ESRC archive contents will remain broadly intelligible to future researchers, without much archive-specific effort to define Representation Information or a Designated Community. In contrast to this simplicity, however, the ESRC archives have to cope with a broad range of associated contextualising metadata, which is different for different projects, and inconsistently or incompletely specified by the originating researchers, perhaps as an afterthought. This makes archive ingest a complicated problem, in contrast to the big science cases, where archive ingest fundamentally involves little more than copying a self-contained set of artefacts from working storage to some preservation store. In particular, the ESRC projects have a complicated set of anxieties about copyright, IPR, confidentiality, anonymization and consent; while LIGO cares intricately about data access and security, it does so in the rather formal context of professional ethics rather than family secrets.

This illustrates two further notable differences between physical science data and that of social science or broader archival resources.

Firstly, the responsibility for ESRC data in practice lies with more junior researchers, helped by part-funded archivists [34, §§5.2.1 & 5.4]. For big science projects, it is funders and senior collaboration members who drive the preservation efforts.

Secondly, essentially all physics data is born digital and complete, meaning that all of the information to be archived is present at the time of deposit. Of course, this is not complete from the point of view of reproducibility (that requires journal articles and personal knowledge) and does not discount the subsequent addition of subjective metadata as finding aids, but it is completely specified from the point of view of conventional future analysis. The distinction is that experimental data is a complete and objective account of everything that was believed to be relevant in recording a physical event which happened at a specific time. One can disagree with the experimenters' beliefs about completeness (this shades into questions of reproducibility and tacit knowledge), complain that some details might be recorded in notebooks rather than digital records (more true of lab-scale than facility-scale experiments), or in extreme cases argue about the nature of objectivity, but a natural science experiment has a much clearer boundary, in space, time and documentary extent, and so a more natural expectation of documentary completeness, than will be usual for an experiment in the social or human sciences. This is different from the traditional archive problem, where the problems of interpretation are more visible and acknowledged, and the problem of incompleteness more evident.

For a vivid and illuminating discussion of the complications and physicality of reproducing experiments, see [35] and references therein (by coincidence, this describes observations amongst gravitational wave experimenters in Glasgow); that discussion is reprised in a larger context in [6, ch.35]. The question of tacit knowledge is discussed at length in [36]. For a discussion of different types of reuse, see [37, §3].

The summary is not that the ESRC or the big science archives have an easier job overall, but that the complications express themselves in different parts of the mapping from OAIS abstractions to local fact. Big science archives must preserve

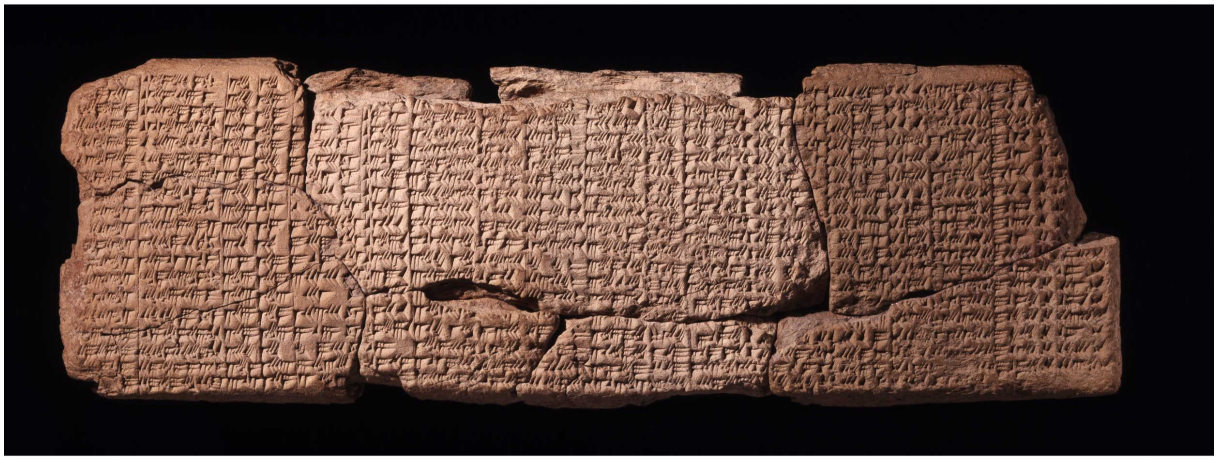

Figure 2: Calculated ephemeris for the period 104 BCE March 23 to 101 BCE April 18, written on Seleucid year 209, month IX, day 18 (103 BCE December 20?). Comparison with a JPL ephemeris shows that the text conjunction times remain within a couple of hours of the correct values, with an offset attributable to an error in the initial value. For detailed discussion, see [40]. British Museum item Sp-II.52, ©Trustees of the British Museum.

large complicated objects for a hard-to-describe Designated Community, but because they are essentially always project-specific archives, their implementation does not have to be generic, and many of the ingestion issues can be baked into the original archive design.

1.8 Babylonian data management (less contrast than you'd think)

Contemporary astronomy began, in the west, in Mesopotamia in the fifth and fourth centuries BCE. Although earlier datasets exist – the *Venus tablet of Ammisaduqa* is a cluster of 7th C BCE copies of 17th C or 16th C data recording the rise times of Venus over a 21 year period – these earlier omen texts seem to have been preserved for largely cultural reasons.

See [38] for background and further references, and [39, ch.4] for very detailed discussion of the physical tablets. The precise date of the observations is of considerable scholarly interest, since an agreed date would provide an absolute fix for the otherwise relative chronology of the Late Bronze Age Near East.

Distinct from these, there is a large set of 4–500 other texts, ranging from 4th C BCE to 75 CE with a smattering going back as far as mid-8th C BCE, and spanning the development of Babylonian theoretical astronomy during the 4th C BCE. These are a mixture of observations, calculated ephemerides (such as Fig. 2), and telegraphically obscure technical documentation. The observation texts – 'astronomical diaries', forming the majority of the texts – describe in sequence celestial and meterological observations, daily commodity prices, river levels, and topical events. The observations of the Sun, Moon and planets were of good enough quality, and preserved over a long enough time, that when babylonian mathematical models were fitted to them they produced values for the synodic and anomalistic months and (implicitly) the orbital periods of the planets, which are very respectably close to their currently-determined values (out by a factor of 3×10^{-7}, in the case of the synodic month). These were used to predict the first and last appearances of planets, and the times of lunar (but not solar) eclipses.

The information in these texts is sometimes available on multiple tablets, although it is not clear whether these duplicates were backups, mirrors, or media refreshes. Many tablets have acquisition metadata, added in ink by the archives, millennia apart, in Babylon and Bloomsbury.

It is clear that the tablets that have survived represent only a small fraction of the total. but both the data, and the mathematical technology that reduced the data and generated the ephemerides, were available and fully intelligible to

Hipparchus (*c* 150 BCE) and, either via him or directly, to Ptolemy (*c* 150 CE). The Babylon Data Centre was still active in the first century CE, though funding cuts meant new acquisitions were by then minimal, and it was operating in the collapsing ruins of the desert city.

The Content Information in the texts is sufficiently well preserved that if the texts can be dated at all (in some cases through contemporary ingest metadata), they can generally be dated to the very day; the technical Representation Information, in contrast, is so terse as to make sense only after the procedure being documented is reconstructed from the Content. The cuneiform presents a challenge, but once this has been transliterated, the datasets are fundamentally intelligible to current astronomers. The preservation strategy is a daring one: by effectively founding western astronomy, and arranging that the data was preserved just long enough that it could be taken over by the (hellenic) successor civilisation, the babylonians ensured that their coordinate system (based on the zodiac) and number system (with angles in degrees, subdivided into base-60 fractions) would still be in use by astronomers 25 centuries later.

This can be classed as a 'high-risk' data preservation strategy, and is not included amongst this report's Recommendations to STFC.

1.9 Bibliographic repositories

Though it is not strictly data, it seems useful to make parenthetical mention of the big science communities' literature repositories, since they seem to illustrate the way in which the communities have learned to act collectively.

The preprint archive at `arXiv.org` started in 1991 as an electronic version of the long-established practice of distributing preprints of accepted journal articles around the high-energy physics community, by post. It currently receives around 6000 submissions per month, predominantly in HEP, astronomy, condensed matter physics and mathematics; it probably receives copies of nearly 100% of the HEP community's output. Authors most typically submit papers at the point when they have been accepted by the journal, but some submit earlier versions, and a few are not further published at all. Although the journals are still providing an *imprimatur*, many papers are now principally read as preprints, and many journals permit citations by arXiv reference. ArXiv is supported by requesting contributions from its heaviest institutional users, on a sliding scale rising to $4 000/year. JISC Collections is one of these 'tier 1' supporters, on behalf of UK colleges and universities.

`http://arxiv.org/Stats/hcamonthly.html`

`http://arxiv.org/help/support/whitepaper`

The NASA ADS at the Smithsonian Astrophysical Observatory preserves bibliographic information for the astronomy literature, holds references to or copies of journal article full texts, and curates digitised copies of older articles sometimes unavailable from publishers. It also curates links between these publications and the arXiv, and between publications and data. See [41] for context, and some discussion of the arXiv numbers mentioned above.

The publication paradigm represented by arXiv (and similar smaller-scale efforts) is underpinned by the peer review processes of journals. However as journal subscription costs rise, journals are progressively cancelled, in a process which may ultimately damage the reviewing process on which the paradigm depends. The SCOAP^3 consortium aims to break out of this cycle by directly supporting a small number of HEP journals, through a levy on the funding agencies which support the field, in proportion to the share of HEP publishing they support. In return for this the journals will remove both subscription charges and page charges for these journals.

`http://scoap3.org/about.html`

1.10 Virtual Observatories

A Virtual Observatory is an astronomical data-sharing system, composed of a network of archives and data-access protocols. The goal is that the data appears to be integrated and ideally appears to be local.

The earliest VOs were Astrogrid in the UK, the US-VO in the US (which became NVO and then VAO), and the Astrophysical Virtual Observatory in Europe (which became Euro-VO). They, along with a growing collection of smaller national or regional VOs, formed the IVOA in 2002. The IVOA exists to broker portable network protocols for sharing data, on the part of cooperating archives, and accessing it, on the part of client applications. The IVOA focuses primarily on 'traditional' astronomy, and so has poor coverage of solar physics and more broadly geophysics (and certainly provides no access to GW data).

http://www.astrogrid.org, http://www.usvao.org/, and http://www.euro-vo.org; plus http://www.ivoa.net.

From this has grown the more general notion of the 'VxO', which is "[a] service that ensures that all resources from sub-field x are known, discoverable, and easily accessible. It looks to the user like a uniform data provider, but it is virtual." Examples include the Virtual Solar-Terrestrial Observatory [42], HELIO, and NASA's Heliophysics Data Environment.

See further commentary in http://lwsde.gsfc.nasa.gov/VxO_Report_Decadal_Survey_5_2011.pdf

http://www.helio-vo.eu and http://lwsde.gsfc.nasa.gov/

1.11 Data products and proprietary periods: reifying data management and release

A common feature of the various data styles above is the notion of the *data product*, and it seems useful to recap and stress the salient features of this here.

> *Data products:* A data product is a designed and documented output of an instrument, intended to be both archivable and immediately useful to other researchers, by virtue of having observational artefacts removed as much as possible.

Depending on the discipline and the engineering complexity of the instrument, data products may be anything from the raw data to a highly processed derivative of the raw data; the ideal data product contains all the scientifically relevant information with none of the experimental artefacts.

Researchers are not restricted to using only data products, but it will only rarely be necessary for them to resort to reanalysing raw data (see the discussion on p.10).

Data products correspond closely to the 'Information Packages' of the OAIS model (see Sect. 3.1.1). In our experience, there tends to be little practical difference between Submission Information Packages (SIPs) and Archival Information Packages (AIPs), and where there are distinct Dissemination Information Packages (DIPs), they tend to be available in addition to the available SIPs and AIPs. An exception to this is archives such as the Wide-Field Astronomy Unit at Edinburgh, which specialises in astronomical survey science, and develops enhanced archives (which is to say, value-added AIPs) as part of its participation in collaborative astronomy projects.

http://www.roe.ac.uk/ifa/wfau/

When the various Packages differ, they tend to be regarded as successively higher-level, as opposed to alternative, data products.

The notion of data products has a number of concrete advantages.

- Most immediately, the existence of a stable and documented output makes it easier for researchers to use and repurpose experimental and observational results.
- Because the products are so central to an instrument's output, they, and the pipelines that produce them, are designed and costed at early stages of an instrument's production.
- Researchers can produce and share software which processes well-defined products, possibly from more than one instrument.
- Because they are so explicit, they form well-defined start and end points of discussions about interoperability between instruments. Indeed, the VO

programme could be characterised as an extended effort to negotiate new common products which archives and software developers agree can be successfully generated (by archives) from existing AIPs.

There is of course a cost associated with the design and development of data products, but we believe that this will in most cases be much smaller than the costs associated with the retrospective documentation and distribution of *ad hoc* datasets.

Another notion that is well-known in the physical sciences, but which as far as we are aware is rare outside, is that of explicit *proprietary periods* for data.

> *Proprietary period:* A 'proprietary period' is a period after data is acquired, and therefore archived, by a shared instrument, during which it is private to the observer or observers who requested it, and after which the data (usually automatically) becomes public.

The term 'embargo period' would possibly be more generally intelligible, but 'proprietary' is conventional. The notion is discussed elsewhere in this document (see for example Sect. 1.4), but we stress it here because it usefully concretizes a number of otherwise vague questions about data release.

Instead of rather broad questions of the how, when, why and whether of data management and release, we instead have questions such as 'what are the data products?', 'whom are they documented for, and how expensively?', 'how long is the proprietary period?' or 'what is the quid pro quo for this period?' These questions don't magically become easy to answer, but they become a lot easier to ask, and invite concrete answers and negotiation rather than *ad hoc* argument.

Compare the comments about Herschel data in Sect. 1.4.

There is nothing in the notions of data products and proprietary periods which is obviously specific to the physical sciences. The notions have become well-established in this area probably because it has long experience, of necessity, of using large shared instruments which are operated to a greater or lesser extent as services. This is less often the case in disciplines with more bench-scale experimental norms, but even some areas of biology are now more often using shared facilities, and in other disciplines, data products and proprietary periods would become more natural, the more that preservation-aware storage is used [43].

We commend the notions of data products and proprietary periods, and the data culture they engender, to the broader research community. Indeed, we recommend that **data managers should consider adopting the language of data products and explicit proprietary periods when designing and documenting their holdings**.

Recommendation 1

2 The responsibilities for data preservation

2.1 Visualising benefits

Why do funders wish to preserve data? Because they perceive *benefits* to that preservation.

Building on this truism, it seems useful to explicitly articulate these benefits. The JISC-funded project Keeping Research Data Safe (KRDS) (see `http://www.beagrie.com/krds.php` and [44]) described a collection of studies and tools supporting data preservation. Amongst the KRDS innovations was a typology of *benefits*, describing three dimensions: direct to indirect, near- to long-term and public to private. In a slight extension to the work in KRDS, we can take the notion of 'dimensions' perfectly literally, assign any particular benefit to a position along each of the three axes, and plot the result in a three-dimensional space; see Fig. 3 on the next page.

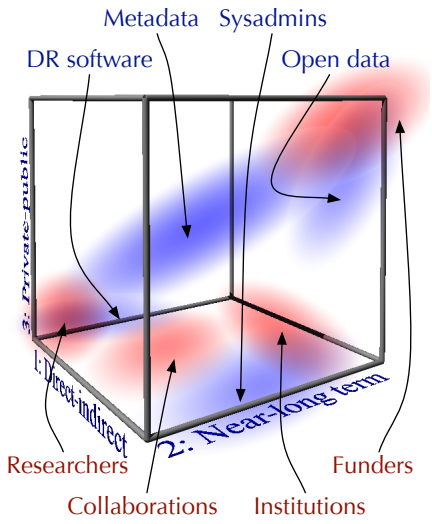

Figure 3: Visualizing benefits

In this figure we identify four benefits which might be associated with a big-science project – namely the existence of data-reduction software, good metadata, the provision of open data and the existence of system administrators – and we sketch the approximate volumes they might occupy along the three axes (in blue). On the same diagram, we can indicate (in red) the approximate areas of interest of four sample stakeholders.

In the example here, 'sysadmin support' can be seen as an indirect benefit to researchers, typically private to an institution, but creating value in the near- and long term; it is therefore spread along the 'near-long term' axis, but at one extreme of the other two dimensions. We can put on the same diagram the approximate areas of interest of various research stakeholders. For simplicity, we are here conceiving of individual researchers as selfish and short-termist, though the same researchers will have long-term interest when they have a collaboration or institutional hat on, and indirect public interests in the long-term health of their discipline when they are serving on a funding council grants panel; below we will take the term 'funders' to refer both to the officials of funding bodies, acting as proxies for the wider interests of society, and to members of the research community discharging service roles.

We should not take this diagram too literally – it is not clear that the axes are independent, and the extent and even the gross positions of the various interests and benefits are debatable. The diagram is nonetheless thought-provoking. For example, it visually predicts that much of the research community is not particularly interested in 'open data' and only incompletely interested in 'good metadata' (in-collaboration researchers care when a dataset was acquired, because they need that information to perform their analyses, but they have little interest in dissemination and licensing metadata, for example, because that is the long-term concern of funders and their proxies). We can therefore naturally conceive of the funders taking the role of the conscience of a discipline, worrying about long-term imponderables so that individual researchers don't have to. It follows from this, that the open data case made to funders, for example, will be an institutionally self-interested one, but that the case made to researchers must be qualitatively different, and be either pragmatic ('you must care because your funders care') or high-minded ('your socio-cultural duty is…'). Neither of these is a poor argument, nor indeed a cynical one, but we are acknowledging here that, to a busy and distracted researcher, the self-interest argument in isolation may have little purchase.

(unless it's *other people's* open data, of course)

2.2 The case for open data

Internationally, there is a push towards such data sharing in the more general context of scholarly research (see for example [45] or [46]). The most explicit statement here is in the NSF's GC-1 document [47], which in section 41 states that "[NSF] expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable." This is reiterated in almost the same words in their 2010 data sharing policy [3]. They additionally require a brief statement, attached to proposals, of how the proposal would conform to NSF's data-sharing policy.

STFC, in common with the other UK research councils, requires that "the full text of any articles resulting from the grant that are published in journals or conference proceedings […] must be deposited, at the earliest opportunity, in an appropriate e-print repository"; it has not yet made any corresponding statement

http://www.scitech.ac.uk/rgh/rghDisplay2.aspx?m=s&s=64

on data releases.

The year 2009 saw some excitement (relating to the incident inevitably labelled 'climategate', and to some other data-release disputes) related to the management and release of climate data. This illustrated the political and social significance of some science data sets; the contrast between what scientists know, and the public believes, to be normal scientific practice; and some of the issues involved in the generation, ownership, use and publication of data. The cases during that year illustrate a number of complications involved in data releases.

http://www.guardian.co.uk/environment/2010/apr/20/climate-sceptic-wins-data-victory

UEA's Climate Research Unit is a partner in the ACRID project, also funded by the JISC MRD programme: http://www.cru.uea.ac.uk/cru/projects/acrid/

1. Data is often passed from researchers or groups directly to others, across borders, with no general permission to distribute it further.

2. Data collection may be onerous, and the result of significant professional and personal investments.

3. Raw data is generally useless without the more or less significant processing which cleans it of artefacts and makes it useful for further analysis.

4. However not all disciplines have the clear notion of published data products which is found in astronomy and which is implicit in the OAIS notion of archival deposit.

5. Science is a complicated social process.

The last point is simultaneously obvious and deeply intricate. Unpacking it would distract us here, but there is further discussion, in a very apposite historical context, in [6], elaborated in [36].

In science, we preserve data so that we can make it available later. This is on the grounds that scientific data should generally be universally available, partly because it is usually publicly paid for, but also because the public display of corroborating evidence has been part of science ever since the modern notion of science began to emerge in the 17th century (CE) – witness the Royal Society's motto, 'nullius in verba', which the Society glosses as 'take nobody's word for it'. Of course, the practice is not quite as simple as the principle, and a host of issues, ranging across the technical, political, social and personal, complicate the social, evidential and moral arguments for general data release.

The arguments *against* general data releases are practical ones: data releases are not free, and may have significant financial and effort costs (cf Sect. 3.4). Many of these costs come from (preparation for) data preservation, since it is formally archived data products that are the most naturally releasable objects: releasing raw or low-level data *may* be cheap, but may also have little value, since raw underdocumented datasets are likely to be useless; or more pessimistically they may have a negative value, if they end up fostering misunderstandings which are time-consuming to counter (this point obviously has particular relevance to politicised areas such as climate science). In consequence of this, the 'open data question' overlaps with the question of data preservation – if the various costs and sensitivities of data preservation are satisfactorily handled, then a significant subset of the practical problems with open data release will promptly disappear. We discuss the data preservation question below, in Sect. 2.3.

It seems worth noting, in passing, that the physical sciences broadly perform better here than other disciplines, both in the technical maturity of the existing archives and in the community's willingness to allocate the time and money to see this done effectively.

What all this indicates is that there is a need for an explicit framework for discussing the pragmatics of open data (cf point 4 above). We can go further and suggest (it is almost a Recommendation) that the OAIS model's notion of an AIP, and its reflection in the notion of a *data product* should be central to this discussion.

2.3 The case for data preservation

The case for data preservation in astronomy was implicitly made in Sect. 1.4: as an observational science, much astronomy data is repeatable, but there are important cases where what is being observed is a slow secular change, or some unpredictable (usually ultimately explosive) event; sometimes data can be opportunistically reanalysed to extract information distinct from the information the observation was designed for. Astronomical data is potentially useful *and* usable almost indefinitely. Thus there is a reasonable expectation that the data can be and will be exploited by unknown astronomers, far into the future.

HEP data is somewhat different (as noted in Sect. 1.5). As an experimental science, it is generally very much in control of what it observes, and is able to design experiments of considerable ingenuity, in order to make measurements of exquisite discriminatory power. A consequence of this is firstly that HEP experiments have a much stronger tendency to become obsolete with each technological generation, and secondly that the complication of the apparatus makes it hard to communicate into the future a level of understanding sufficient to make plausible use of the data. Experimental apparatus will generally be understood better and better as time goes on (this is also true of satellite-borne detectors in astronomy), so that data gathered early in an experiment will be periodically reanalysed with increased accuracy. However this understanding is generally not preserved formally, but is pragmatically communicated through wikis, workshops, word of mouth, configuration and calibration files, and internal and external reports. Even if all of the tangible records were magically preserved with complete fidelity, and supposing that the more formal records do contain all the information required to analyse the raw data, an archive would still be missing the word-of-mouth information which a new postgrad student (for example) has to acquire before they can understand the more complete documentation. We can think of this as a 'bootstrap problem'. In OAIS terms, the Representation Network for HEP data is particularly intricate, and while the Representation Information nearest to the Data Object may be complete, it may be infeasible to gather the Representation Information necessary to let a naive researcher make sense of it. The Designated Community for HEP data may therefore be null in the long term.

This sounds pessimistic, but [26] describes a number of scenarios in which HEP data can and should be reanalysed some decades after an experiment has finished, and describes ongoing work on the development of consensus models for preserving data for long enough to enable such post-experiment exploitation. This provides a strong case for a style of preservation somewhat different from the astronomical one. What these models have in common is a commitment of staff to actively conserve and continuously exploit the data. This post-experiment staff can therefore be conceived as a form of walking Representation Information so that, while they are still involved, the data might have a Designated Community which corresponds to those individuals in a position to undertake an extended apprenticeship in the data analysis (this model is further discussed on p.32).

GW data is, as usual, somewhere between these two extremes. As astronomy, the GW data consists of unrepeatable measurements which will potentially be of value to astronomers well into the future; as a HEP-style experiment it makes those measurements using two or three generations of highly sophisticated apparatus, each generation of which will improve on the sensitivity of its predecessors by orders of magnitude. An additional feature, however, is that no-one has ever convincingly detected a gravitational wave, though there have been repeated claims of detection in the past, so that the first claims by LIGO or aLIGO will be scrutinized particularly closely.

Finally, and as noted in Sect. 2.2, if data is well archived, then most of the pragmatic objections to opening that data do not apply. Thus, to the extent that

general data release is a good in itself, it is a further argument in favour of a well supported archive.

2.4 Should raw data be preserved?

In the data-preservation world, there is often an automatic expectation that 'everything should be preserved', so that an experiment can be redone, results reanalysed, or an analysis repeated, later. Is this actually true? Or if it is at least desirable, how much effort should be expended to make it true? This question is implicit in, for example, the discussion of software preservation in Sect. 3.2.

When a physical experiment is set up and working, it is usual to avoid tinkering with it as much as possible, to avoid any unexpectedly significant change. That is, even with a small-scale lab-bench experiment, it is accepted that not everything can be effectively documented, and that an experiment might not be immediately replicable purely from published information (cf [6, ch.35] and Sect. 1.7). This expectation (or rather, lack of expectation) is also true of larger-scale experiments, which might be financially, professionally or, at the largest scales, politically infeasible to replicate. Perhaps this attitude should extend to other aspects of the experimental process.

It is because very large-scale experiments are impossible to replicate, and even hard for an external reviewer of an article to criticise meaningfully, that large collaborations submit their publications to extremely scrupulous internal review. See `http://stuver.blogspot.com/2011/03/big-dog-in-envelope.html` for a post-mortem account of such a review.

In many cases, the pipeline for reducing raw data seems to fall into this category: it encodes hard-to-document information, but is itself hard to document, hard to use, and unlikely ever to be reused in fact. If this software is not preserved, then the raw data is effectively unreadable, which means there is no case for preserving it. There is therefore a case that at least some details of the experimental environment – digital as well as physical – are not reasonably preservable, and that as a result little effort should be expended on preserving them.

It is data products that make raw data less necessary. It is feasible to document the scientific meaning of data products, and the community expects that a project will provide this documentation as part of the publication of the products (indeed, it the documentation that makes these *products* rather than just a casual data snapshot). The data products allow researchers to dig beneath the conclusions of a particular article (or indeed the contents of a higher-level data product), and to criticise and build on what they find there. Higher-level products are the result of higher-level scientific judgements, and it is normal for these to be regenerated by researchers other than the originators, either using their own software or the originators' pipelines. These later-stage pipelines are more formally supported by projects, which involves making them reasonably portable, so that they are both easier to preserve as well as being more valuable objects of preservation.

We should stress that we are not advocating deliberately deleting raw data, and its associated pipelines – it *might* be useful, and it *might* be usable – but simply noting that one should not overstate its value.

2.5 OAIS: suitability and motivation

In Sect. 3.1.1, we provide an overview of the OAIS model, and describe how it relates to astronomical data.

The OAIS standard is formally a product of the Consultative Committee for Space Data Systems (CCSDS), and with this in its lineage it is quite naturally matched to the data management problems of the physical sciences. Essentially all the explicit and implicit assumptions of the OAIS standard are true in the area we are studying: the data producer (a satellite or a detector) is usually obvious, the various Information Packages (or data products) well understood, and the Designated Community easily identified.

The motivation for a digital preservation standard, as discussed in the OAIS standard itself [4, §2], is that digital preservation represents a double problem:

`http://www.ccsds.org`

There is no contradiction here with the remarks in Sect. 1.7 about the difficulty of describing the Designated Community of science archive users. It is easy to name a science Designated Community, but it may be hard to describe ahead of time what those community members can be expected to know. A social science archive may have an unpredictably broad range of ultimate users, but using the archive will need little specialist knowledge; in contrast a particle physics dataset will probably be of interest only to particle physicists, but the normal education of such a physicist three decades hence, and thus the content and extent of the specialised Representation Information that Community will need, might be very hard to guess at.

(i) digital information is intrinsically harder to preserve than traditional information, which is capable of sitting on a shelf in a well-understood and intelligible format, and mouldering at a well-understood and graceful rate; and (ii) more and more organisations are producing digital information *and* are implicitly expected to archive their own material. This means that these non-specialist archives have a complicated task to perform, which is potentially at odds with the daily urgencies of their main business.

This *appears* to mean in turn (and in JISC contexts it is often taken to mean in practice) that these organisations need as much detailed and prescriptive help as possible, ideally devolving their archive responsibilities to a central discipline- or funder-specific archive, to the extent possible while respecting the low-level complications and friction alluded to in Sect. 1.7. This is not the model which is appropriate for big-science datasets.

2.6 What should big-science funders require, or provide?

We have described several common features of big-science data management in Sect. 1.3. and we have outlined some particular contrasts with other communities in Sect. 1.7. As noted in Sect. 0.1, our focus here is on STFC's strategically funded projects, rather than the smaller projects funded by individual research grants.

Big-science data sets are generally intimately coupled to solutions to leading-edge technical challenges, and cannot usefully be regarded as incremental changes to previous solutions. This, coupled with the general availability of extensive technical expertise within such communities, means that any generic solution is very unlikely to be appropriate, and that it is both reasonable and feasible to require custom archiving solutions for such projects. There is no *recipe* for data preservation on this scale, and all that can be hoped for is a structured approach to a custom solution. Having said this, not even the most innovative science experiments are so completely *sui generis* that they warrant a data preservation approach which is reimagined from scratch. It is therefore wasteful to ignore the considerable intellectual investments in the OAIS model, the growing penumbra of commentaries on and developments of it, and the minor industry of validation and auditing efforts related to it.

Recommendation 2

We are therefore led to the conclusion that the most effective overall strategy for effective data management in the large-scale experimental physical sciences is that **funders should simply require that a project develop a high-level DMP plan as a suitable profile of the OAIS specification [4]**. This profile should be detailed enough to require negotiation with the funder and with the experiment's community, but can leave many of the implementation details to the good engineering judgement of the project's management. We believe the LIGO DMP plan [5] can be taken to be exemplary in this regard.

Recommendation 3

Big-science projects have the technical skills, the management structures, and the budgets to take on such a task, and to deliver a custom archive which can be shown to meet identified goals. We recommend that **funders should support projects in creating per-project OAIS profiles which are appropriate to the project and meet funders' strategic priorities and responsibilities**.

The discussion in Sect. 3.5 suggests that one result of the development of an OAIS-based DMP plan is that the resulting plan is explicit enough to generate useful deliverables, and to benefit from the growing interest in OAIS 'validation'.

We suggest the following specific funder actions.

- Actively engage with projects to help them develop an OAIS profile. This will include overview literature, including the OAIS specification, tutorial reports such as [48], and commentary such as [49], or perhaps specialised

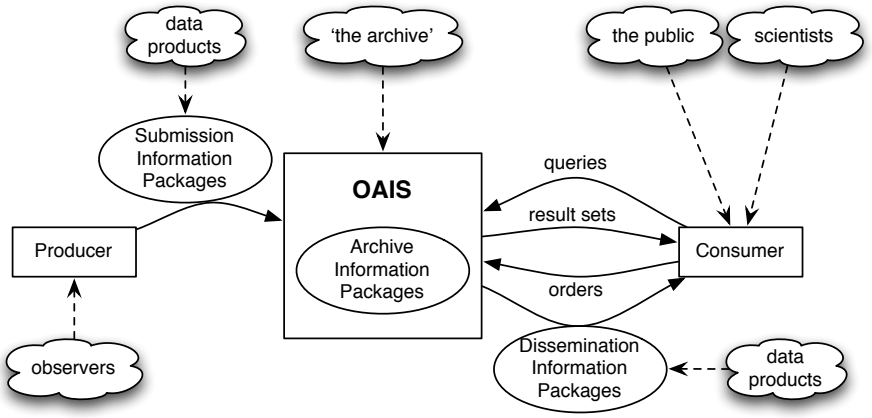

Figure 4: The highest-level structure of an OAIS archive, annotated with the corresponding labels from conventional astronomical practice (redrawn from [4, Fig. 2-4]). The dissemination data products will typically be the same as the submitted ones, but archives can sometimes create value-added ones of their own.

workshops if necessary. These are high-level introductions, rather than procedure-based tick-lists.

- Develop or support expertise in criticising and validating such OAIS profiles. For example, the CASPAR consortium (see for example [50]) has developed strategies for detailed validation of projects' claims about long-term data migration. Similar work – for example validating a project's assumptions about its Designated Community – would reassure the wider community that the archive design is likely to achieve its goals for the future.

The first of these is reasonably straightforward, consisting of little more than gathering resources. The second is a longer-term project which may require some expertise to be built up and supported at a funder-supported facility (such as RAL, in the UK), or through liaison with the DCC.

A corollary of this more active engagement is that funders must financially support the preservation work they require. See Sect. 3.4.

3 The practicalities of data preservation

3.1 Modelling the archive

3.1.1 The OAIS model

We introduce here the main concepts of the OAIS model. Full details are in [4] with a useful introductory guide in [48] and some discussion in the LSC context in [5]; the OAIS motivation is further discussed in Sect. 2.5.

The term *OAIS* stands for an *Open Archival Information System*. The word 'open' is not intended to imply that the archived data is freely available (though it may be), but instead that the process of defining and developing the system is an open one. The principal concern of an OAIS is to preserve the usability of digital artefacts for a pragmatically defined long term. An OAIS is not only concerned with storing the lowest-level *bits* of a digital object (though this part of its concern, and is not a trivial problem), but with storing enough *information* about the object, and defining an adequately specified and documented *process* for migrating those bits from system to system over time, that the information or knowledge those bits represent can be retrieved from them at some

indeterminate future time. The OAIS model can therefore be seen as addressing an administrative and managerial problem, rather than an exclusively technical one.

The OAIS specification's principal output is the *OAIS reference model*, which is an explicit (but still rather abstract) set of concepts and interdependencies which is believed to exhibit the properties that the standard asserts are important (Fig. 4 on the preceding page). The OAIS model can be criticised for being so high-level that "almost any system capable of storing and retrieving data can make a plausible case that it satisfies the OAIS conformance requirements" [49], and there exist both efforts to define more detailed requirements [49], and efforts to devise more stringent and more auditable assessments of an OAIS's actual ability to be appropriately responsive to technology change [50].

An OAIS archive is conceived as an entity which preserves objects (digital or physical) in the Long Term, where the 'Long Term' is defined as being long enough to be subject to technological change. The archive accepts objects along with enough Representation Information to describe how the digital information in the object should be interpreted so as to extract the information within it (for example, the FITS specification is Representation Information for a FITS file). That Information may need further context – for example, to say that a file is an ASCII file requires one to define what ASCII means – and the collection of such explanations turns into a Representation Network. This information is all submitted to the archive in the form of a SIP agreed in some more or less formal contract between the archive and its data producers.

Once the information is in the archive, the long-term responsibility for its preservation is *transferred* from the provider to the archive, which must therefore have an explicit plan for how it intends to discharge this.

The archive distributes its wares to Consumers in one or more Designated Communities, by transforming them, if necessary, into the DIP which corresponds to a 'data product'. The members of the Designated Community are those users, in the future, whom the archive is designed to support. This design requires including, in the AIP, Representation Information at a level which allows the Designated Community to interpret the data products *without ever having met one of the data Producers*, who are assumed to have died, retired, or forgotten their email addresses.

The OAIS model originated within the space science community, so it can be mapped to the physical science data of the GW community without much violence.

3.1.2 The DCC Curation Lifecycle model

The OAIS model is on the face of it a linear one, and suggests that data is created, then ingested, then preserved, and then accessed, in a process which has a clear beginning and end. This is compatible with the observation that one point of archiving data is to reuse or repurpose it, creating new archivable data products in turn, but this longer-term cycle remains only implicit in the model. The OAIS model is therefore very usefully explicit about those aspects of archival work concerned with long-term preservation, but its conceptual repertoire is such that a discussion framed by it runs the risk of underemphasizing the range of roles a data repository has, or even of marginalising it.

We thank Dorothea Salo, of the University of Wisconsin library, for emphasizing to us the useful applicability of the DCC model to the case of big science data, and Angus Whyte, for elaborating the contrasts between the DCC and OAIS models.

In contrast, the DCC has produced a lifecycle model [51] (Fig. 5 on the next page) which stresses that data creation, management, and reuse are part of a cycle in which preservation planning, for example, can naturally happen before data creation as well as after it; and in which data can be appraised, reappraised, and possibly disposed of if it becomes obsolete. It therefore makes explicit both the short- and long-term cycles in the flow of active research data, and it emphasizes the active involvement of data curators in maintaining that cycle.

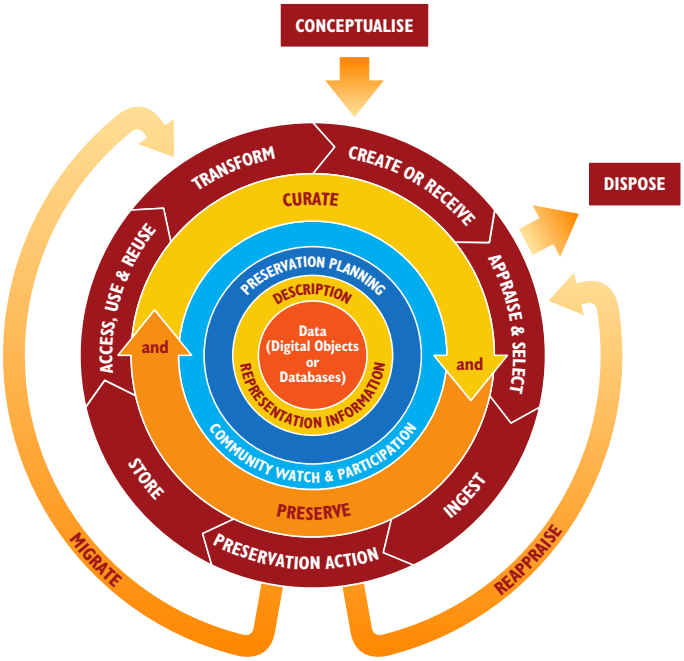

Figure 5: The DCC lifecycle model, from [51]

Cycles of use and re-use are not the only links between datasets. As discussed in [52], one digital object can also provide context for another, in a variety of ways. To some extent this remark rediscovers the notion of the OAIS Representation Network, and this in turn prompts us to stress that although we have contrasted OAIS and DCC here, they are not in competition: OAIS is concerned with the creation and management of a working archive with gatekeepers and firm goals; the DCC model is concerned with the location of the archive in the wider intellectual context.

The DCC model is immediately compatible with the observation, in Sect. 3.4 below, that HEP and GW archives effectively avoid some preservation costs by seeing long-term preservation as only part of the role of a data repository. Accepting data, making it available as working storage, transforming it into immediately useful forms, or appraising (possibly regenerable) datasets whose storage costs outweigh their usefulness, all give the archive a familiarity with the data, and the researchers a familiarity with the archive, which means that the decision to select certain data for long-term preservation is potentially more easily reached, more easily defended and more easily funded, than if the archive is conceived as a cost-centre bucket bolted on the side of the project. This appears to be borne out by the LIGO experience, in which the new DMP plan was developed and successfully argued for by the same personnel who were long responsible for the design and management of the data management system on which everyone's daily work depends.

3.2 Software preservation

As discussed in, for example, Sect. 1.6.2, there is often a substantial amount of important information encoded in ways which are only effectively documented in software, or software configuration information. There is therefore an obvious case for preserving this software (though note the caveats of Sect. 2.4).

Preservation of a software pipeline requires preserving the pipeline software itself, a possibly large collection of libraries the software depends on, the oper-

ating system (OS) it all runs on, and the configuration and start-up instructions for setting the whole thing in motion. The OS may require particular hardware (CPUs or GPUs), the software may be qualified for a very small range of OSs and library versions, and it may be hard to gather all of the configuration information required (there is some discussion of how one approaches this problem in for example [26]). It is not certain that it is necessary, however: if the data products are well-enough described, then re-running the analysis pipeline may be unnecessary, or at least have a sufficiently small payoff to be not worth the considerable investment required for the software preservation. We feel that, of the two options – preserve the software, or document the data products – the latter will generally be both cheaper and more reliable as a way of carrying the experiment's information content into the future, and that this tradeoff is more in favour of data preservation as we consider longer-term preservation.

This last point, about the changing tradeoff, emphasizes that the two options are not exclusive: one can preserve data *and* preserve software, and the JISC-funded Software Sustainability Institute provides a growing set of resources which provide guidance here. However the solutions presented generally focus on active curation, in the sense of preserving software through continuing use and maintenance. This can be successful, and is the approach implicit in [26], but it seems brittle in the face of significant funding gaps, and would not deal well with the case where a software release is deliberately unused, for example because it has been superseded.

http://www.software.ac.uk/

The UK Starlink project provided astronomical software. It ran from 1980 to 2005, when it was rescued from oblivion by being taken up by the UK Joint Astronomy Centre Hawai'i. The current distribution includes still-working code from the 80s. The Netlib and BLAS libraries have components which date from the 70s.

3.3 Data management planning

3.3.1 DMP in space

As one might expect, both NASA and ESA have formalised DMP plans.

NASA's National Space Science Data Center (NSSDC) has led NASA's data planning since the mid-80s. It was initially the NSSDC which negotiated a Project DMP plan with missions, but since the 1990s this has become the responsibility of the NASA Planetary Data System (PDS). The NSSDC's data retention policy describes what categories of data product should be retained indefinitely, and the PDS provides resources to mission planners on the processes and tools for preparing data for preservation.

ESA's Planetary Science Archive "provides expert consultancy to all of the data producers throughout the archiving process. As soon as an instrument is selected, PSA begin working with the instrument team to define a set of data products and data set structures that will be suitable for ingestion into the long-term archive." The ESA archive is by design compatible with the PDS.

http://nssdc.gsfc.nasa.gov

http://pds.nasa.gov/

http://nssdc.gsfc.nasa.gov/nssdc/data_retention.html

http://pds.nasa.gov/tools/index.shtml

We are grateful to Paul Butterworth of NASA for helpful advice here.

http://www.rssd.esa.int/index.php?project=PSA&page=about

3.3.2 Current and future DMP in the LSC

The current LIGO DMP plan [5], discusses DM planning with an emphasis on the preparations for the eventual public data release.

The LIGO DMP plan proposes a two-phase data release scheme, to come into play when aLIGO is commissioned; this was prepared at the request of the NSF, developed during 2010–11, and will be reviewed yearly.

The plan documents the way in which the consortium will make LIGO data open to the broader research community, rather than (as at present) only those who are members of the LSC. This document describes the plans for the data release and its proprietary periods, and outlines the design, function, scope and estimated costs of the eventual LIGO archive, as an instance of an OAIS model. This is a high-level plan, with much of the detailed implementation planning delegated to partner institutions in the medium term.

In the first phase, data is released much as it is at present: validated data will be released when it is associated with detections, or when it is related to papers announcing *non*-detections (for example, associated with another astronomical event which might be expected or hoped to produce detectable GWs). In the second phase – after detections have become routine, and the LIGO equipment is acting as an observatory rather than a physics experiment – the data will be routinely released in full: "the entire body of gravitational wave data, corrected for instrumental idiosyncrasies and environmental perturbations, will be released to the broader research community. In addition, LIGO will begin to release near-real-time alerts to interested observatories as soon as LIGO *may* have detected a signal." This second phase will begin after LIGO has probed a given volume of space-time (see [5, ref 7]), *or* after 3.5 years have elapsed since the formal LIGO commissioning, whichever is earlier. Alternatively, LIGO may elect to start phase two sooner, if the detection rate is higher than expected.

In phase two, the data will have a 24-month proprietary period.

The DMP plan describes three (OAIS) Designated Communities. Quoting from [5, §1.5], the communities are as follows.

- LSC scientists: who are assumed to understand, or be responsible for, all the complex details of the LIGO data stream.
- External scientists: who are expected to understand general concepts, such as space-time coordinates, Fourier transforms and time-frequency plots, and have knowledge of programming and scientific data analysis. Many of these will be astronomers, but also include, for example, those interested in LIGO's environmental monitoring data.
- General public: the archive targeted to the general public, will require minimal science knowledge and little more computational expertise than how to use a web browser. We will also recommend or build tools to read LIGO data files into other applications.

The LIGO DMP plan is, we believe, a good example of a plan for a project of LIGO's size: it is specific where necessary, it was negotiated with the project's funder (NSF) so that it achieved their goals, and it went through enough iterations with the broader LIGO community (the agreed version in [5] is version 14) that its authors could be confident it had their approval, and that the community was comfortable with what the DMP plan was proposing. The document has a strong focus on the LIGO data release criteria, since this was the most immediate concern of both the funder and the project, but it systematically lays out a high-level framework for future data preservation, guided by the OAIS functional model.

3.4 Data preservation costs

There is a good deal of detailed information, and some modelling, of the costs of digital preservation. The KRDS2 study [44, §§6&7] includes detailed costings from a number of running digital preservation projects, in some cases down to the level of costings spreadsheets. The LIFE^3 project has also developed predictive costings tools [53], and the PLANETS project (`http://www.planets-project.eu/`) has generated a broad range of materials on preservation planning, including costing studies.

Although there is a broad range of preservation projects surveyed in the KRDS report, there are numerous common features. Staff costs dominate hardware costs, and scale only very weakly with archive size. The study also notes that acquisition and ingest costs are a substantial fraction (70–80%) of overall staff costs, but also scale very weakly with archive size. These are relatively small archives, generally below a few TB in size, where ingest is a significant

component of the workload. In this report we are interested in archives three or four orders of magnitude larger than this where (as discussed below) ingest may be cheaper, but in broad terms, it appears still to be true that staff costs dominate hardware costs at larger scales, and scale only weakly with archive size.

Parenthetically, notice that the above discussion prompts the question 'what is the size of an archive?' The number of bytes it consumes is an obvious and readily available measure, but may not be particularly meaningful in this context. The number of items (such as interview transcripts, images or database rows) may be a better measure, and still objectively identifiable, archive by archive. If there were some measure of abstract information content, we speculate that this is what would scale most straightforwardly with the effort required for quality control and metadata curation, and hence with staff effort. We hesitate to ask what such a measure might be, in case the answer is 'citation analysis'.

The lack of scaling with size, even when an archive progressively grows in size, seems to suggest that it is an archive's *initial* size (in the sense of small, medium or large, for the time) that largely governs the costs.

We were given access to confidential figures for the development and operations of a mid-to-large size astronomy archive (of order 10 TB of relational data and 100 TB of flat file data), developed by an experienced archive site. The archive software and system development cost 25–30 staff-years of effort: the bulk of this was for the core database system, but between a quarter and a third was for software to support ingest and the generation of data products. The organisation budgets around 3 FTEs for operation of this archive, which includes ingest, quality control and helpdesk support (this is an estimated fraction of an operations team covering several archives at the same site, so there may be some economies of scale). About a quarter of the annual operating budget is spent on hardware.

We are most grateful to Fernando Comerón, of ESO, for sharing these figures.

The European Southern Observatory (ESO) data archive manages data from multiple ESO facilities; it shares space with the still-developing ALMA archive, but the figures below do not include ALMA. The archive is based on spinning disks backed by a tape library (for further details, see [54]). It currently holds 190 TB, increasing at around 7 TB month^{-1}. The hardware costs average around 330 k€ yr^{-1}, which includes hardware replacement and data migration, and which has remained flat for some years, despite the slowly increasing data volumes. Running costs amount to 55 k€ yr^{-1} (some smaller systems account for part of this), and licences, networks and other consumables account for about 30 k€ yr^{-1}. Manpower costs come to 4 FTEs of ESO staff plus around 270 k€ yr^{-1} of outsourced staff. Neither hardware nor software costs appear to scale with data volume, with some cost elements even dropping as the archive moves to completely on-line data distribution.

There is some discussion of the CDS funding model in Sect. 1.4.1.

http://pds.nasa.gov/tools/cost-analysis-tool.shtml

The NASA PDS has developed a parameterized model for helping proposers estimate the costs involved in preparing data for archiving in the PDS; most relevantly for the above discussion it includes a scaling with data volume of $1 + 1.5 \log_{10}(\text{volume/MB})$ (that is, a multiplier which increases by 1.5 for each order of magnitude increase in data volume).

As noted in Sect. 1.5, the HEP community is now constructing more detailed plans for data preservation, and the associated costs. Reference [26] estimates that a formal long-term archive (a level-3 or -4 archive, in the terms of that paper) would cost 2–3 FTEs for 2–3 years after the end of the experiment, followed by 0.5–1.0 FTE/year/experiment spent on the archive's preservation. They compare this to the 100s of FTEs spent on for the running of the experiment, and on this basis claim an archival staff investment of 1% of the peak staff investment, to obtain a 5–10% increase in output (the latter figure is based on their estimate that around 5–10% of the papers resulting from an experiment appear in the years immediately after the experiment finishes; since this latter figure is derived on the

current model, which achieves this without any formal preservation mechanisms, this estimate of the return on investment in archives may be optimistic).

It is worth noting that in astronomical, HEP and GW contexts, archive ingest is generally tightly integrated with the system for day-to-day data management, in the sense that data goes directly to the archive on acquisition and is retrieved from that archive by researchers, as part of normal operations. On the other side of the archive, projects will generate and disseminate data products – which look very much like OAIS DIPs – as part of their interaction with external collaborators, without regarding these as specifically archival objects. Thus the submissions into the archive may consist of both raw data and things which look very much like DIPs, and the objects disseminated will include either or both very raw and highly processed data. The *long-term* planning represented in the LIGO DMP plan [5], for example, is therefore less concerned with setting up an archive, than with the adjustments and formalizations required to make an existing data-management system robust for the archival long term, and more accessible to a wider constituency. What this means, in turn, is that some fraction of the OAIS ingest and dissemination costs (associated with quality control and metadata, for example) will be covered by normal operations, with the result that the *marginal* costs of the additional activity, namely long-term archival ingest and dissemination, are probably both rather low and typically borne by infrastructure budgets rather than requiring extra effort from researchers. This is corroborated by our informants above, who generally regard archive costs as coming under a different heading from 'data processing costs'. The point here is not that the OAIS model does not fit well – it fits very well indeed – nor that ingest and dissemination do not have costs, but that if the associated activities can be contrived to overlap with normal operations, then the costs directly associated with the archive may be significantly decreased. This is the intuition behind the recent developments in 'archive-ready' or 'preservation-aware storage' (cf [43] and Sect. 3.1.2), and confirms that it is a viable and effective approach.

This is consistent with the ERIM project's conclusions that "ideally information management interventions should result in a zero net resource increase" [55, p.8]. In this case there is no extra resource required from the researchers, though there might be a need for extra resource under an infrastructure heading.

As a final point, we note that big-science projects are inevitably also large-scale engineering projects, so that the consortia and their funders are broadly familiar with the procedures, uncertainties and management of cost estimates, so that the costing and management of data preservation can be naturally built in to the relationship between funders and funded, if the funders so require it.

As is shown by the vagueness of some of the remarks above (despite sometimes very specific numbers), there seems little in the way of a consensus model for the costing of the long-term preservation of large-scale data. There will surely be detailed costings for the management of PB-scale data for commercial organisations, but these are not likely to be useful for our purposes, since they are more concerned with immediate business continuity than multi-decade archives, are serving different technical communities, and are likely to be extremely confidential.

We therefore recommend that **STFC should develop a costings model for the publication and preservation of data, which is matched to the data challenges of the big-science community**. We expect that this can build on the domain-agnostic work already done in this area by JISC, and on the detailed work done on closely related problems by NASA's cost-estimation community [56].

Recommendation 4

3.5 The GW community and the AIDA toolkit

The AIDA Self-Assessment Toolkit [57] is a (JISC funded) set of qualitative benchmarks for discussing at how developed an institution's archive is. It leads an archive manager through a set of a few dozen elements, inviting them to grade their archive from 1 (poor) to 5 (international exemplar). The goal is not to produce a pass/fail score, but instead to help archive managers understand their current and future requirements, and to "enable an institution to decide whether

The AIDA document links these five stages, rather alarmingly, to a five-step programme developed at Cornell, which starts with acknowledging that you have a problem, and goes, via institutionalisation, to "embracing […] dependencies", noting that "you can't do it alone". Clearly, data-management planning is habit-forming.

specific actions need to be taken in regard to particular assets, or when and how it is desirable to improve on its current capabilities". The AIDA authors acknowledge that the assessment is simplistic and subjective, but stress that "AIDA aims to allow you to evaluate your institution against a recognised capability scale, and then suggests appropriate actions based on that evaluation". The AIDA goal is to model the progress of an archive from the acknowledgement that an archive is desirable, through to the exemplary externalisation of the archive as a resource.

In Appx. B, we list our estimates of the scores for LSC data management. We hope these assessments are of specific use to the GW community, but believe that the discussion in general may be of use to other, similarly structured, big science communities.

The scores for the current LSC cluster in the middle, around three (which corresponds to 'consolidate' in the Cornell model). This is an impressive score for a project which is, from one point of view, doing only what is regarded as normal for a well-run large-scale physics experiment. The higher scores are generally associated with the formality and auditability of the long-term plans, rather than any qualitatively different practice, and we believe that these scores will naturally drift upwards as a result of the development of an explicit DMP plan, structured using the OAIS concept set, in collaboration with a suitably critical funder.

The toolkit is broken into organisational, technology and resources (generally funding) 'legs'.

The 'organisational leg' is concerned with the high-level support for the archive. To the extent that it is meaningful, the average for these scores is above three (which is good). The lower scores are generally associated with the informality of the current archive (compared to a service-oriented commercial organisation) rather than any more concrete inadequacy: the data is backed up and reasonably findable, though this reflects cultural norms within the physical sciences rather than something a particular archiving plan can take credit for.

The 'technology leg' is concerned with the hardware and personnel support for data management. As with the organisational leg, the GW community scores highly here without really trying, simply because the community has long experience of managing *and sharing* large volumes of data. The lower scores are again associated with the current informality of operations (from the point of view of an archive as opposed to a working data-management infrastructure), and these will naturally rise when the LSC's DMP plan is implemented and reviewed.

The scores in the 'resources leg' are the least well-justified. The LSC generally scores well, in the sense that we can be confident that there will be resources to support an archive effort – it's seen as a high-importance activity – even though there are few resources currently explicitly earmarked for this. This section may therefore be useful for suggesting what budget lines should eventually exist.

4 Conclusions and recommendations

In this report, we have described some of the ways in which 'big science' manages its data, as part of a broader data culture which is characterised by large collaborations, and which has decades of experience in agreeing how, and when, and when not, to share data.

We can say with some confidence that the big science data culture manages its data well (and this seems to be corroborated by the AIDA assessment discussed in Sect. 3.5), but we are not suggesting that other disciplines could or should simply copy this culture, since there are various reasons (cf, Sect. 1.3) why this culture is particularly natural in some areas.

There are however some practices which we do believe are straightforwardly portable to other disciplines. As we discuss in Sect. 1.11, the notions of *data products* and *proprietary periods* very naturally concretize otherwise diffuse ar-

guments about data management and sharing, transforming them from 'whether' and 'why' to 'which' and 'how long'. As well, we believe that embedding data management in the day-to-day practice of researchers lowers costs in both the short term (researchers can easily re-find their own data, and interpret others') and the long term (since preservation becomes a technical problem of conserving an in-use repository). We discuss the costing of data management at slightly greater length in Sect. 3.4.

We repeat our explicit recommendations below.

1. Data managers should consider adopting the language of data products and explicit proprietary periods when designing and documenting their holdings (Sect. 1.11, p21).
2. Funders should simply require that a project develop a high-level DMP plan as a suitable profile of the OAIS specification [4] (Sect. 2.6, p26).
3. Funders should support projects in creating per-project OAIS profiles which are appropriate to the project and meet funders' strategic priorities and responsibilities (Sect. 2.6, p26).
4. STFC should develop a costings model for the publication and preservation of data, which is matched to the data challenges of the big-science community (Sect. 3.4, p33).

Acknowledgements

This project was funded by JISC, as part of the 'Managing Research Data' programme. We are most grateful to the numerous people who have commented on various drafts of this report, or provided us with information or resources. In particular, we thank Stuart Anderson (LIGO), Paul Butterworth (NASA), Harry Collins (Cardiff), Fernando Comerón (ESO), Joy Davidson (DCC), Françoise Genova (CDS), Magdalena Getler (DCC), Simon Hodson (JISC), Sarah Jones (DCC), Dorothea Salo (Wisconsin), Angus Whyte (DCC), and Roy Williams (LIGO).

A Case study

We have produced a detailed discussion of the structure of the LIGO working data management system, as a separate document [29]. This document is currently available only within LIGO: those observations which have not been incorporated into this present report are probably too detailed to be of general interest. We hope, however, that the case-study will be of some use internally to to the LSC.

B AIDA assessment

The AIDA self-assessment toolkit [57] is a JISC-funded set of qualitative benchmarks for assessing how developed an institution's archive is. See Sect. 3.5 for discussion.

The labels in the table below are sometimes a little cryptic; refer to the full toolkit for useful elaborations.

The answers below generally refer to the *early 2011* state of the LSC archive arrangements, on the grounds that concrete answers to a variant question are preferable to speculative answers to a future one. These are probably a reasonable indication of the likely status of a forthcoming formal archive, but in a few case, as noted, we can give no meaningful answer.

In the scores below, level 1 is 'poor', and level 5 is 'international examplar'.

Organisational leg

1: institution-wide mission statements (5) The LIGO project has prepared a formal DMP, at funder request

2: institutional policies for asset management (3) LIGO has prepared a formal DMP, and is addressing political and cultural reservations, awaiting funding and implementation

3: review mechanisms at Institutional level (4) As well as the DMP, there already exist well-understood collaboration-wide review procedures, and these will be used to review the plan on an annual basis

4: institutional capability for sharing assets (3) Current storage is, of necessity, distributed; the collaboration manages this informally but effectively, however this is generally working storage, and not regarded as archival storage

5: institutional level of contingency planning (3) There is no formal centralised asset management. Continuity is regarded as a technical matter which can reasonably be left to the professional good practice of the sites managing the distributed storage. As before, this is currently regarded as working rather than archival storage.

6: institutional capability for audit (3) Extensive logs exist, but are not centralised nor in any standard format; files, once created, are not expected to be modified, though there is no way to verify that this is true in fact

7: institutional monitoring mechanisms (4.5) All data and processes are open to the entire collaboration, and most processes are widely discussed; the collaboration is its own user-base. There are (by design) no external users of the data, nor yet any external review of the mechanisms.

8: extent of institutional conformance to metadata management (2 to 4) Metadata is devised in a somewhat *ad hoc* way by individual instruments or software elements (stage 2), but this is also added and managed thoroughly, and in accordance with what is regarded as experimental good practice (stage 4)

9: extent of institutional contracts (3) Not applicable to current working storage

10: institutional understanding of IPR (5) Formal MoUs between partners regarding access to data, and clear guidance from funders regarding the eventual release of the data

11: institutional disaster planning (2) As with asset continuity, this is currently regarded as a technical matter for storage managers

Technological leg

1: institutional infrastructure (5) The collaboration has considerable technical resource, and interoperates well. Planning is informal but effective. The sophisticated user-base is comfortable with this informality, but this could in principle become a liability when the resource management moves from a development to a service model.

2: appropriateness of institutional technologies (4) There is plenty of appropriate technology, though the plan for the archival management of assets is not yet detailed

3: integrity of institutional backup and storage (3) Important data is backed up (possibly by mirroring), as part of normal operations

4: institutional processes (2) Uncertain: what there is will be done as part of normal operations

5: institutional understanding of obsolescence (3.5) High general awareness, and occasional discussion, but at present little formal planning

6: institutional capability (4) Changes to processes are widely and frankly discussed, and documented as internal publications; change is managed effectively, but relatively informally

7: institutional capability for security (3) There is a high level of awareness of the need to keep the data proprietary, but given the scientific context, there are no likely attack scenarios as such; the problem will largely evaporate once the data is finally released publicly

8: institutional security mechanisms (3.5) No formal threat analyses, but the security is probably appropriate to the level of threat; day-to-day attacks (ie not specifically targeted at this data) are the responsibility of distributed storage and computing managers

9: institutional disaster plan and capacity for business recovery (3) Not applicable to the current experimental phase

10: institutional capacity to create metadata (4) Almost all metadata is added automatically (compare organisational.08)

11: effectiveness of an Institution-wide repository (2) LIGO has prepared a formal DMP

Resources leg

1: institutional business planning (2) LIGO is preparing a formal DMP

2: institutional capacity for review (4) DMP to be reviewed annually; project as a whole has close relationships with funders and stakeholders

3: institutional capability for resource allocation (4) Resource planning is coordinated at a senior level

4: institutional capability for risk management (2) General awareness at present, but this should become clearer in future DMP iterations

5: institutional business transparency (4.5) Depending on the precise meaning intended, this could be 4 or 5. There is substantial auditing from collaboration funders

6: institutional capacity for sustainable funding (3.5) Good relationships with funders mean that funding is probably predictable on five- to ten-year timescales, but unpredictable in the longer term. However the main funder (NSF) has expressed a strategic commitment to long-term data preservation.

7: institutional staff management (3) Not applicable to the current experimental phase

8: institutional management of staff numbers (3) Not applicable to the current experimental phase

9: institutional commitment to staff development (3) Not applicable to the current experimental phase

About this document

LIGO-P1000188-v10, 2011 June 29 First public version, available at https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=p1000188

v1.1, 2012 July 14 Minor revisions, some added material, and typos and minor errors corrected. There are a couple of additional sections, but no changes to section or figure numbers. The pagination will have changed in places.

References

[1] Norman Gray and Graham Woan. Digital preservation and astronomy: Lessons for funders and the funded. In I. N. Evans, A. Accomazzi, D. J. Mink, and A. H. Rots, editors, *Proceedings of ADASS XX*, volume 442 of *ASP Conference Series*, pages 13–16. ASP, 2011. Available from: http://www.aspbooks.org/publications/442/013.pdf, arXiv:1103.2318.

[2] Simon Hodson. Managing research data programme [online]. Available from: http://www.jisc.ac.uk/whatwedo/programmes/mrd [cited 14 May 2010].

[3] Dissemination and sharing of research results [online]. March 2011. Available from: http://www.nsf.gov/bfa/dias/policy/dmp.jsp [cited 10 May 2011].

[4] Reference model for an open archival information system (OAIS) – CCSDS 650.0-B-1. CCSDS Recommendation, 2002. Identical to ISO 14721:2003. Available from: http://public.ccsds.org/publications/archive/650x0b1.pdf.

[5] Stuart Anderson and Roy Williams. LIGO data management plan. LIGO Technical Report, 2011. Available from: https://dcc.ligo.org/public/0009/M1000066/014/LIGO-M1000066-v14.pdf.

[6] Harry M Collins. *Gravity's shadow: the search for gravitational waves*. University of Chicago Press, 2004.

[7] Harry M Collins. LIGO becomes big science. *Historical Studies in the Physical and Biological Sciences*, 33:261–297, 2003. doi:10.1525/hsps.2003.33.2.261.

[8] Brian Moe. LIGO data grid: Data set size estimates [online]. Available from: https://www.lsc-group.phys.uwm.edu/lscdatagrid/resources/data/sizes.html [cited 20 May, 2010].

[9] A. R. Taylor. The square kilometre array. *Proceedings of the International Astronomical Union*, 3(Symposium S248):164–169, 2007. doi:10.1017/S1743921308018954.

[10] Cisco. Entering the zettabyte era. White Paper, July 2011. Available from: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.pdf.

[11] High Level Expert Group on Scientific Data. Riding the wave. Final report, European Commission, October 2010. Available from: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

[12] Alex Ball. Review of the state of the art of the digital curation of research data. Project report erim1rep091103ab12, University of Bath, 2010. Available from: http://opus.bath.ac.uk/19022/.

[13] PARSE.Insight Project. Deliverable d3.3: Case studies report. Project deliverable, 2010. Available from: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf.

[14] N. Hambly, H. MacGillivray, M. Read, S. Tritton, E. Thomson, B. Kelly, D. Morgan, R. Smith, S. Driver, J. Williamson, Q. Parker, M. Hawkins, P. Williams, and A. Lawrence. The SuperCOSMOS sky survey – i. introduction and description. *Monthly Notices of the Royal Astronomical Society*, 326(4):1279–1294, 2001. arXiv:astro-ph/0108286, doi:10.1111/j.1365-2966.2001.04660.x.

[15] Derek Jones. The scientific value of the Carte du Ciel. *Astronomy & Geophysics*, 41(5):16–21, 2000. doi:10.1046/j.1468-4004.2000.41516.x.

[16] Yudhijit Bhattacharjee. Stars in dusty filing cabinets. *Science*, 324(5926):460–461, April 2009. doi:10.1126/science.324_460.

[17] F. R. Stephenson and L. V. Morrison. Long-term fluctuations in the earth's rotation: 700 BC to AD 1990. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 351(1695):165–202, 1995. Available from: http://rsta.royalsocietypublishing.org/content/351/1695/165.abstract, doi:10.1098/rsta.1995.0028.

[18] L. Jetsu, S. Porceddu, J. Lyytinen, P. Kajatkari, J. Lehtinen, T. Markkanen, and J. Toivari-Viitala. Did the ancient egyptians record the period of the eclipsing binary Algol – the Raging one? Preprint, April 2012. To appear in Astron. Astrophys. arXiv:1204.6206.

[19] G. J. Toomer. A survey of the Toledan Tables. *Osiris*, 15:pp. 5–174, 1968. Available from: http://www.jstor.org/stable/301687.

[20] FITS Working Group. Definition of the flexible image transport system (FITS). Technical report, Commission 5 of the International Astronomical Union, 2008. Available from: http://fits.gsfc.nasa.gov/fits_standard.html.

[21] William D Pence, L. Chiappetti, Clive G Page, R. A. Shaw, and E. Stobie. Definition of the Flexible Image Transport System (FITS), version 3.0. *Astronomy and Astrophysics*, 524:A42+, December 2010. doi:10.1051/0004-6361/201015362.

[22] European Space Agency. The Hipparcos and Tycho catalogues. Online, 1997. ESA SP-1200. Available from: http://www.rssd.esa.int/index.php?project=HIPPARCOS.

[23] F. Genova, D. Egret, O. Bienaymé, F. Bonnarel, P. Dubois, P. Fernique, G. Jasniewicz, S. Lesteven, R. Monier, F. Ochsenbein, and M. Wenger. The CDS information hub. *Astron. Astrophys. Suppl. Ser.*, 143(1):1–7, 2000. doi:10.1051/aas:2000333.

[24] F. Ochsenbein, P. Bauer, and J. Marcout. The VizieR database of astronomical catalogues. *Astron. Astrophys. Suppl. Ser.*, 143(1):23–32, 2000. doi:10.1051/aas:2000169.

[25] Andrew Curry. Rescue of old data offers lesson for particle physicists. *Science*, 331(6018):694–695, February 2011. doi:10.1126/science.331.6018.694.

[26] David M South. Data preservation in high energy physics. In *Proceedings of the 18th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2010)*, January 2011. arXiv:1101.3186.

[27] Max Boisot, Markus Nordberg, Saïd Yami, and Bertrand Nicquevert, editors. *Collisions and Collaboration*. Oxford Scholarship Online, 2011. doi:10.1093/acprof:oso/9780199567928.001.0001.

[28] Avgousta Kyriakidou-Zacharoudiou and Will Venters. Distributed large-scale systems development: Exploring the collaborative development of the particle physics grid. In *5th International conference on e-Social Science, Cologne*, 2009. Available from: http://ssrn.com/abstract=2109945.

[29] Tobia D Carozzi, Norman Gray, and Graham Woan. LIGO data: a case-study in big science databases. Technical case study T1000368, LSC, 2010.

[30] Matthew Pitkin, Stuart Reid, Sheila Rowan, and Jim Hough. Gravitational wave detection by interferometry (ground and space). *Living Reviews in Relativity*, 14(5), 2011. Available from: http://relativity.livingreviews.org/Articles/lrr-2011-5/.

[31] LSC and Virgo. Memorandum of understanding between VIRGO and LIGO. LSC Memorandum M060038, LSC and Virgo consortia, 2009.

Version 1, of 2009 March 11. Available from: https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=1156.

[32] LSC. Bylaws of the LIGO scientific collaboration. LSC Memorandum M050172, LSC, 2010. Version 5, of 2010 January 15. Available from: https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=8432.

[33] LSC. LIGO Scientific Collaboration publication and presentation policy. LSC Memorandum T010168, LSC, 2007. Version 3, of 2007 August 8. Available from: https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=26956.

[34] Veerle Van den Eynden, Libby Bishop, Laurence Horton, and Louise Corti. Data management practices in the social sciences. Technical report, UK Data Archive, July 2010. Available from: http://www.data-archive.ac.uk/media/203597/datamanagement_socialsciences.pdf.

[35] Harry M Collins. Tacit knowledge, trust and the Q of sapphire. *Social Studies of Science*, 31(1):71–85, 2001. doi:10.1177/030631201031001004.

[36] Harry M Collins and Robert Evans. *Rethinking Expertise*. Chicago University Press, 2007.

[37] Sean Bechhofer, John Ainsworth, Jiten Bhagat, Iain Buchan, Philip Couch, Don Cruickshank, David De Roure, Mark Delderfield, Ian Dunlop, Matthew Gamble, Carole Goble, Danius Michaelides, Paolo Missier, Stuart Owen, David Newman, and Shoaib Sufi. Why linked data is not enough for scientists. In *IEEE International Conference on eScience*, pages 300–307, Los Alamitos, CA, USA, 2010. IEEE Computer Society. doi:10.1109/eScience.2010.21.

[38] Asger Aaboe. Babylonian mathematics, astrology and astronomy. In *The Cambridge Ancient History*, volume 3 (part 2), pages 276–292. Cambridge University Press, 1991. doi:10.1017/CHOL9780521227179.

[39] Russell Hobson. *The Exact Transmission of Texts in the First Millennium BCE: An Examination of the Cuneiform Evidence from Mesopotamia and the Torah Scrolls from the Western Shore of the Dead Sea*. PhD thesis, University of Sydney, 2009. Available from: http://ses.library.usyd.edu.au/bitstream/2123/5404/1/r-hobson-2009-thesis.pdf.

[40] Asger Aaboe. Scientific astronomy in antiquity. *Phil. Trans. R. Soc. Lond.*, A276:21–42, 1974. doi:10.1098/rsta.1974.0007.

[41] Alberto Accomazzi. The role of repositories and journals in the astronomy research lifecycle. Slides from talk at Astroinformatics 2010, Pasadena, June 2010. Available from: http://www.astroinformatics2010.org/pdfs/Accomazzi.pdf.

[42] Peter Fox, Deborah L. McGuinness, Luca Cinquini, Patrick West, Jose Garcia, James L. Benedict, and Don Middleton. Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience. *Computers & Geosciences*, 35(4):724–738, 2009. doi:10.1016/j.cageo.2007.12.019.

[43] Michael Factor, Dalit Naor, Simona Rabinovici-Cohen, Leeat Ramati, Petra Reshef, and Julian Satran. The need for preservation aware storage. *ACM SIGOPS Operating Systems Review*, 41(1):19–23, 2007. Available from: http://www.research.ibm.com/haifa/projects/storage/datastores/papers/preservation_data_store_osr07_dec_30.pdf, doi:10.1145/1228291.1228298.

[44] Neil Beagrie, Brian Lavoie, and Matthew Woollard. Keeping research data safe 2. JISC Project Report, April 2010. Available from: http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf.

[45] Peter Arzberger, Peter Schroeder, Anne Beaulieu, Geof Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhlir, and Paul Wouters. Science and government: An international framework to promote access to data. *Science*, 303(5665):1777–1778, 2004. Available from: http://www.sciencemag.org/cgi/reprint/303/5665/1777.pdf, doi:10.1126/science.1095958.

[46] Raivo Ruusalepp. Infrastructure planning and data curation: A comparative study of international approaches to enabling the sharing of research data. Technical report, Digital Curation Centre, November 2008. Available from: http://www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf.

[47] National Science Foundation. Grant general conditions (GC-1). Technical Report gc1010, National Science Foundation, 2010. Available from: http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gc1010.

[48] Brian F Lavoie. The open archival information system reference model: Introductory guide. DPC Technology Watch Series Report 04-01, OCLC, January 2004. Available from: http://www.dpconline.org/docs/lavoie_OAIS.pdf.

[49] David S H Rosenthal, Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito. Requirements for digital preservation systems: A bottom-up approach. *D-Lib Magazine*, 11(11), 2005. Available from: http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html.

[50] CASPAR Consortium. CASPAR D4104: Validation/evaluation report. FP6 project deliverable, October 2009. Available from: http://www.casparpreserves.eu/Members/cclrc/Deliverables/caspar-validation-evaluation-report/at_download/file.

[51] Digital Curation Centre. DCC curation lifecycle model [online]. 2010. Available from: http://www.dcc.ac.uk/resources/curation-lifecycle-model [cited 22 June 2011].

[52] Christopher E Lee. Taking context seriously: A framework for contextual information in digital collections. Technical Report SILS Technical Report 2007-04, University of North Carolina, October 2007. Available from: http://sils.unc.edu/sites/default/files/general/research/TR_2007_04.pdf.

[53] Brian Hole, Li Lin, Patrick McCann, and Paul Wheatley. LIFE^3: A predictive costing tool for digital collections. In *iPres*, September 2010. Submitted to iPres. Available from: http://www.life.ac.uk/3/docs/Ipres2010_life3_submitted.pdf.

[54] Paul Eglitis and Dieter Suchar. Historical lessons, inter-disciplinary comparison, and their application to the future evolution of the ESO archive facility and archive services. In *Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*. European Space Agency, December 2009. Available from: http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/34_Eglitis_ESO_ARCHIVE.pdf.

[55] Mansur Darlington, Alex Ball, Tom Howard, Steve Culley, and Chris McMahon. RAID associative tool requirements specification (version 1.0). Technical Report ERIM Project document erim6rep101111mjd10, University of Bath, 2011. To appear. Available from: http://opus.bath.ac.uk/22811.

[56] NASA Cost Analysis Division. Cost estimating handbook (2008 edition). Online, 2008. See also http://cost.jsc.nasa.gov/. Available from: http://www.nasa.gov/offices/pae/organization/cost_analysis_division.html.

[57] University of London Computer Centre. The AIDA self-assessment toolkit mark II, February 2009. Available from: http://aida.jiscinvolve.org/toolkit.

Glossary

Terms marked 'OAIS' are copied from the OAIS specification [4, §1.7.2].

ADS Astrophysical Data Service: a bibliographic archive for astronomy, based at the Harvard-Smithsonian Center for Astrophysics; ADS preserves full-text copies of journal articles, both in collaboration with publishers, and through a digitization process, and maintains a widely-used bibliographic ID system (`http://ads.harvard.edu`). 10, 11, 19

AIP Archival Information Package: An Information Package, consisting of the Content Information and the associated Preservation Description Information, which is preserved within an OAIS (OAIS). 20, 23, 28

aLIGO Advanced LIGO: The successor project to LIGO, due to start in 2015. 14, 15, 24, 30

arXiv A electronic preprint service, see `http://arxiv.org`. The arXiv started in the early 90s, based on FTP and email. It initially serviced particle physics and astronomy, but has expanded to cover other areas of physics, mathematics, and some areas of computing science. 10

ATLAS One of the four detectors at the LHC, and one of the two large ones. 6, 14

big science A class of science projects characterised by being international, highly collaborative and expensively funded (see Sect. 1.1 for more discussion). 4, 5, 34

catalogue In the astronomical context, a catalogue is a table of positions and other information for stars or other other astronomical objects. 9--11

CCSDS Consultative Committee for Space Data Systems: authors of the OAIS reference model, see `http://www.ccsds.org`. 25

CDS Strasbourg Data Centre: (see `http://cdsweb.u-strasbg.fr/` and Sect. 1.4.1). 11, 32

CMS Compact Muon Solenoid: One of the four detectors at the LHC, and one of the two large ones. 14

Content Information The set of information that is the original target of preservation by the OAIS (OAIS). 17, 19

Data Object Either a Physical Object or a Digital Object (OAIS) (that is, the `Data Object' is the sequence of bits, or the physical object which is *the data* in the most primitive sense). 24

data products Formal data outputs from an observatory, instrument or process (see Sect. 1.4). 10

data sharing The formalised practice of making science data publicly available. 22

DCC Digital Curation Centre: `http://www.dcc.ac.uk` (not to be confused with the LSC Document Control Center). 4, 5, 27, 28

Designated Community An identified group of potential Consumers who should be able to understand a particular set of information (OAIS). 11, 17, 24, 25, 27, 28, 31

DIP Dissemination Information Package: The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS (OAIS). 20, 28, 33

DMP Data Management and Preservation. 12, 16, 26, 29--31, 34

ESA European Space Agency: http://www.esa.int. 6, 11, 12, 30

ESO European Southern Observatory: A pan-european agency running a set of southern-hemisphere telescopes http://www.eso.org. 32

ESRC Economic and Social Research Council: the principal social science funder in the UK, see http://www.esrc.ac.uk. 17

FITS Flexible Image Transport System: the standard file format in astronomy, see http://fits.gsfc.nasa.gov. 10

GEO A German-British consortium, responsible for the GEO600 interferometer, funded jointly by STFC and the German government. 14

GEO600 The GEO observatory located near Hannover in Germany. 14

GW Gravitational Wave. 4, 5, 8, 12, 14

HEP High Energy Physics. 8, 12, 13

HERA A particle accelerator at the German DESY facility. 13

Information Package The Content Information and associated Preservation Description Information which is needed to aid in the preservation of the Content Information. The Information Package has associated Packaging Information used to delimit and identify the Content Information and Preservation Description Information (OAIS). 20, 25

IVOA International Virtual Observatory Alliance: the consortium which defines VO standards. 12, 19

JISC Joint Information Systems Committee: The organisation responsible for the maintenance and effective exploitation of the academic computing network in the UK, and the funders of this present report. 4, 5, 33, 35

KRDS Keeping Research Data Safe: JISC project developing and documenting data preservation tools and studies; see http://www.beagrie.com/krds.php and [44]. 21

LHC The Large Hadron Collider at CERN: the accelerator is the host for two large general purpose detectors (ATLAS and CMS) and two smaller ones (ALICE and LHCb). 6, 12, 13

LIGO Laser Interferometer Gravitational-wave Observatory: the hardware, comprising LIGO Lab and GEO (see http://ligo.org and Sect. 1.6.1). 5, 6, 12, 14, 16, 17, 30

LIGO Lab The Caltech/MIT consortium, funded by NSF to design and run the LIGO interferometers in the US, see http://www.ligo.caltech.edu. 14

Long Term A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future (OAIS). 8, 28

LSC LIGO Scientific Collaboration: The network of research groups contributing effort to the LIGO experiment and data analysis, see http://ligo.org. 4, 5, 14, 16

LVC A data-sharing agreement between the LSC and the Virgo Collaboration (see Sect. 1.6.1). 5, 14

MOU Memorandum of Understanding: the relationships between the various participating entities and the LSC is articulated through a series of annually reviewed MOUs. 14

MRD Managing Research Data: A funding programme within the JISC e-Research theme, see http://www.jisc.ac.uk/whatwedo/programmes/mrd. 4, 23

NASA National Aeronautics and Space Administration: The US space agency http://www.nasa.gov. 6, 9, 30, 32, 33

NSF National Science Foundation: the principal (non-defence) science funder in the USA. 4, 22, 30

NSSDC National Space Science Data Center: the permanent archive for NASA space science mission data http://nssdc.gsfc.nasa.gov. 30

PDS Planetary Data System: The NASA data archive and standard set http://pds.nasa.gov/. 30, 32

pipeline A software system (or sometimes a software-hardware hybrid) which transforms raw data into more or more levels of data product. The data reduction pipelines, which must be able to keep up with the rate at which data is acquired, and which are assembled from a mixture of standard and custom software components, generally absorb a significant fraction of the total development budget of a new instrument. 11, 14, 15, 29

raw data The data extracted directly from an instrument or observation; since it is uncalibrated and uncorrected, it is generally of little use to those not intimately familiar with the instrument (see Sect. 1.4). 10, 11

Representation Information The information that maps a Data Object into more meaningful concepts (OAIS). 11, 13, 17, 19, 24, 28

Representation Network The set of Representation Information that fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the Long Term (OAIS). 24, 28

SIP Submission Information Package: An Information Package that is delivered by the Producer to the OAIS for use in the construction of one or more AIPs (OAIS). 20, 28

SKA Square Kilometre Array: a low frequency radio telescope with a large (one square kilometre) collecting area. 6

STFC Science and Technology Facilities Council: the primary UK funder of facility-scale science, see http://www.stfc.ac.uk. 4, 5, 22

strain data The fundamental GW signal. 15

Virgo Italian-French gravitational-wave detector http://www.virgo.infn.it/. 14

VizieR A repository of astronomical catalogue data at CDS (see Sect. 1.4.1). 10, 11

VO Virtual Observatory: a set of data sharing argreements and protocols. See Sect. 1.10 (not to be confused with grid Virtual Organisations). 5, 19, 20

Index

See also the Glossary, which additionally serves as the index of acronyms