Digital Preservation and Astronomy: Lessons for funders and the funded

Norman Gray, Graham Woan and Tobia Carozzi
University of Glasgow, UK
LIGO–G1000747-v2
What is ‘big science’?
What is ‘long-term’ (and for whom?)
What is OAIS? (and why?)
Preserving and opening data
What is the data preservation problem (and whose?)
What is the data preservation solution?
big science

big money – decades, G€ / G$
big author lists – LIGO=0.8 kAuth; ATLAS=3 kAuth
big data – aLIGO ~ 1PB/yr; ATLAS ~ 10 PB/yr (= ‘1 LHC’)
big admin – MOUs, councils, ...
big careers – PhD to tenure

JISC study restricted to this scale of projects – unique to physical sciences
Babylonian data can be used for earth slowdown studies

Plates are used for some astrometry

Astronomers can (roughly) read 1627 Rudolphine tables

...and with help, 12C Toledan tables

So, let’s say a millennium

Venus tablet of Ammisaduqa is a 7th C BCE copy of 17th C data - the rise times of Venus over a 21 year period. This is digital data. Very long-term preservation is possible.
Major challenge to understand data
...so software is crucial
...and supercession inevitable
So, perhaps 30 years?
Features of both astronomy and HEP

No detection announced so far, but still ~PB/yr

Data reduction heavily dependent on software

...but the eventual data products will be intelligible
So ‘long-term preservation’ means answering the question:

Who cares?
And are they born, yet?
And will they speak english, mandarin, or klingon?
And how many legs will they have?

Depending on how long–term your long term is, there’s a variety of entertaining questions and answers
But not necessarily how to make progress
Alternatively, and more pragmatically:

The ‘long term’ means having to change media at least once

(a rather boring answer, admittedly, but rather more amenable to analysis)
OAIS
Open Archival Information System

CCSDS 650.0 = ISO 14721:2003

A high-level model for archives

= ‘a set of terms to think with’

If you’re an archivist, you probably know about this already; if you’re not, it’s a term you should probably at least have heard of.
2.3.1 MANAGEMENT INTERACTION

Management provides the OAIS with its charter and scope. The charter may be developed by the archive, but it is important that Management formally endorse archive activities. The scope determines the breadth of both the Producer and Consumer groups served by the archive.

Some examples of typical interactions between the OAIS and Management include:

– Management is often the primary source of funding for an OAIS and may provide guidelines for resource utilization (personnel, equipment, facilities).

– Management will generally conduct some regular review process to evaluate OAIS performance and progress toward long-term goals.

– Management determines, or at least endorses, pricing policies, as applicable, for OAIS services.

– Management participates in conflict resolution involving Producers, Consumers and OAIS internal administration.

Effective Management should also provide support for the OAIS by establishing procedures that assure OAIS utilization within its sphere of influence. For example, management policies should require that all funded activities within its sphere of influence submit data products to the archive and also adhere to archive standards and procedures.

Broadly intelligible (naturally, since it came from space-data); maps to astro data naturally

The ‘designated communities’ are easy to identify
data preservation planning
We preserve data, so we can make it available later

Data should be available/open!

Because it’s usually publicly paid for

...and that’s how science works

But: precedence, proprietary periods, misunderstandings, audience, documentation, money, time, ...

I’m not pretending it’s easy, but it does have to get done
Physical sciences score well without really trying (yes, the rest of science is _much_ worse)
JISC-funded study of Gravitational Wave community, as proxy for STFC-funded big science

See: http://purl.org/nxg/projects/mrd-gw

Short version: they’re doing the right thing

Not yet much to see at that URL
so what is ‘the right thing’?

- Formal & costed data management planning
- Identification of ‘designated communities’
- Identification of data products (AIPs in OAIS-speak)
- Timescales and criteria for data release
- Framed with OAIS conceptual model
- ...so coupled with the OAIS validation industry

All of this counts as pretty radical stuff, outside the physical sciences, but is easy within. Which is good
so our recommendations will be...

- Big-science funders should say: “read and profile OAIS”
- ...and develop or support the expertise in criticising the result
- ...and use that profile as a framework for validation
- ...and pay for it.

Comments? Yes please! http://purl.org(nxg/projects/mrd-gw and norman@astro.gla.ac.uk}