



# Data preservation & the Virtual Observatory

---

Bob Mann

Wide-Field Astronomy Unit  
Royal Observatory Edinburgh

*rgm@roe.ac.uk*

# Plan

---

- Three basic questions
  - and the implications of the answers to them
- The situation beyond astronomy
- Conclusions



# Conclusions

---

- Data preservation & the VO go hand in hand
- Action needed now
  - Within our own community
  - Interacting with other communities



# 1. Why do we preserve our data?

---

- Because we believe they will be re-used
- Re-use
  - New science – e.g. via integration with other data
  - Checking published results
- Tricky to assess which data will be re-used
  - Easier to say which data *can* be re-used



# Basic conditions for data re-use

---

- Efficient mechanisms for discovering, accessing and analysing relevant data
- Discovery
  - VO: *Resource Metadata, Dataset Characterisation*
- Access
  - VO: suite of protocols developing: SIAP, SSAP,...
- Come back to Analysis later



# Data discovery in the VO

---

- Efficient discovery: rich, accurate & complete metadata that can be queried quickly
  - Accurate, complete: straightforward to prepare
  - Quick to query: stored in simple structure
- How to provide rich content in a simple structure that is straightforward to prepare?
  - Solving this is crucial for VO success

(Moving the metadata from the registry to the data service moves the problem, but doesn't remove it)



# Data Access and Analysis in the VO

---

- Current model (assumed by S\*AP, etc)
  - Astronomer downloads data to own institution to analyse – and that's his/her problem
- Increasing importance of surveys driven by large-scale statistical analyses means this is not sufficient
- Must have data analysis services at the data centre, callable from within VO



## 2. For how long must we preserve our data?

---

- Decades
  - Proper motions, long-term variability, orbits
- Two concerns:
  - Technical: necessity of migration between media
  - Sociological: longer than careers of individual staff and lifetimes of project consortia
- Need bit preservation & logical preservation
  - Should the VO be addressing these?



# 3. Who preserves our data?

---

- Different in different countries
  - Some national data centres – permanent(?)
  - UK: university research groups on rolling grants
- Where does the institutional responsibility for long-term preservation lie?
  - e.g. universities? - associate domain-specific data centres with university libraries?
- Funding agencies find it hard to address long-term issues...but they may have to



# A wider perspective

---

Same issues in many disciplines, leading to

- High-level policy statements
  - e.g. *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007)
- Interdisciplinary research
  - e.g. in UK Digital Curation Centre



# OECD Principles

---

- Openness, Flexibility, Transparency, Legal Conformity, Protection of Intellectual Property, Formal Responsibility, **Quality**, **Professionalism**, **Interoperability**, Security, Efficiency, Accountability, **Sustainability**
- These principles are what our science ministers say will guide their future actions
  - e.g. in UK, new policies from BBSRC and MRC



# Interdisciplinary research

---

- Most work has common starting point
  - Open Archive Information System Reference Model: an abstract model of an archive against which real systems can be assessed
- Influence extends into commercial sector:
  - e.g. IBM White Paper "*Towards OAIS-based Preservation-Aware Storage*"
- We must start taking the OAIS RM seriously



(More on the OAIS RM from Dave Giaretta, no doubt)<sup>12/14</sup>

# Aside: our data aren't all digital

---

- ROE Plate Library:
  - 19,000 plates, only ~1/4 scanned – mainly systematic sky surveys
- Harvard Plate Collection ~500,000 plates
- How much can/should we do with these?
  - What level of access to them can be offered?
  - How well can they be characterised?



# Conclusions

---

- Data preservation & the VO go hand in hand
  - The VO needs the data and can enable the re-use which justifies their preservation
- Action needed now
  - Within our own community
    - Get metadata standards right
    - Analyse what we do in the language of the OAIS RM
  - Interacting with other communities
    - Leverage work based on OAIS RM
    - Enjoy benefits from new high-level data policies

