

# Long Term Archive – Part 1

9/11/03

Tom Kalvelage  
USGS Land Remote Sensing Program  
605.594.6556  
Kalvelage@usgs.gov

# Agenda

---

- **NASA – MOU Status**
- **EOS Data Transition White Paper Review**
  - ◆ **Transition Scenarios**
    - Assumptions
    - Schedule
    - Costs
    - Results
- **Where do we go from here?**
- **What do we do now?**

# Basic Principle of LTA

---

Long term archiving  
*means*  
archiving long term.

# MOU Status

---

- **The 1988 USGS/NASA MOU addressed more than just the LP DAAC.**
- **According to the MOU, NASA agreed to:**
  - ◆ Place the.. long-term archives for land remotely sensed data obtained by NASA at EDC, for “...the EOS program and other current and future experimental systems...”.
  - ◆ Transfer responsibility for active long-term archiving and appropriate science support activities to the USGS.

# MOU Status

---

- **By the MOU, the USGS agreed to:**
  - ◆ Have the EROS Data Center serve as the long-term archive.
  - ◆ Assume responsibility for active long-term archiving and appropriate science support activities for that data in the active short-term archives.

# MOU Status

- **By the MOU, the USGS and NASA agreed to jointly:**
  - ◆ Define the scope and content of the active short- and long-term land remotely sensed archives and associated science support activities covered by the MOU.
  - ◆ Define a budget strategy for the cooperative program which identifies the important complementary roles in earth system science of NASA's EOS and the USGS active long-term archive.
  - ◆ Participate in joint presentations to NASA, DOI, OMB, and the Congress, as necessary to explain the essential roles of each organization and funding needs for the cooperative program.
- **This suggests a high level of cooperation that addresses both short- and long-term archiving, and that strictly speaking has not yet happened.**

# MOU Status

- **A clarification was issued in 1989.**
  - ◆ Transfer of data from the NASA-funded short-term archive to the USGS-funded long term archive was described.
    - Envisioned as a transfer of funding responsibility.
    - NASA would continue to pay 100% of the ingest cost, the USGS would pay an increasing share of the archive and distribution costs starting about three years after data acquisition.
    - The data would not move to a new system.
  - ◆ Knowledge from recent experience raises issues here:
    - Data is constantly being reprocessed, so when the 3 year clock starts is hard to determine.
    - The plans and to some extent the costs of the system the data is currently in are unknown.
    - The new CDR approach may call for a new transfer protocol.

# Is the MOU Still Relevant?

- **Does NASA still want to transfer the data to the USGS?**
  - ◆ NASA does not consider itself an operations agency.
  - ◆ Transferring long-term archive operations to other agencies is consistent with its mission, policy, and intent.
- **Does the USGS still want to accept the data?**
  - ◆ Archiving and using low- and medium-resolution land remote sensing data and derived products is a strategic goal of the USGS, Geography Discipline, and EDC.
  - ◆ The data are applicable and useable for both the USGS and the land science and applications community that the USGS serves.
- **The details of the transfer is still TBD, but both agencies indicate they want it to happen.**

# Transition

- **Transition Scenarios – 2 degrees of freedom**
  - ◆ Data set ownership.
    - Separate data sets (static data sets transferred at one time).
    - Shared data sets (data sets transferred incrementally).
      - ◆ *We will see that reprocessing plays havoc with this.*
  - ◆ Data system ownership.
    - Separate data systems (USGS builds it's own).
    - Shared data system (USGS shares an evolved ECS).
- **First, we will examine the separate data set scenario, with both shared and separate data systems.**
- **As we go, we will point out the differences the shared data set scenario might impose.**

# Assumptions

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Mission Status				Terra Ends		Aqua Ends	Terra transferred		Aqua transferred	
USGS Milestones		Plan	Plan	System Analysis	Design & Dev	Dev & Test	Start Terra Ops	Terra Ops, Dev Aqua	Terra and Aqua Ops	

- **Schedule Assumptions:**

- ◆ Terra and Aqua end on schedule (L+6 years).
- ◆ Three years after the missions end (L+9 years):
  - All NASA reprocessing is done.
    - ◆ In shared or separate data sets, only the final reprocessed version of the data set is archived long term.
    - ◆ Basically, the tradeoff is “X and Y” vs. “2X and no Y”.
  - The data transfers to the USGS.

# Assumptions

- **Data Assumptions**

- ◆ Assumes only one version of each data set is transferred.
- ◆ Assumes all MODIS and ASTER data in the LP DAAC will transfer to the USGS.
  - This is unlikely, see Part 2 of the LTA discussion.
- ◆ Assumes none of the data at any other DAAC or site transfers to the USGS.
  - This is also unlikely, but harder to estimate.
  - In particular, USGS is interested in lower level MODIS data.
- ◆ Assumes no further NASA data from other missions will be transferred to the USGS as part of this estimate.
  - This is still TBD.
- ◆ These assumptions are made to simplify estimate.

# Assumptions

- **Ops and Maintenance (e.g. DAAC) Assumptions.**
  - ◆ Operations and Maintenance costs decreases 4% / year.
    - -4% per year means FY12 budget is 64% of FY03.
    - If possible at all, this requires significant efficiencies, and capacity reductions in everything but ops and user services.
- **Development / Build Assumptions.**
  - ◆ Moore's law is assumed to hold.
  - ◆ If NASA system is used or shared, the USGS would have to build some links to their systems.
    - Assumes NASA gives USGS the ECS hardware for 3-5 yrs.
  - ◆ If the USGS builds a new system, we assume it will be built to a minimal set of requirements, providing a level of service equivalent to the USGS today.
  - ◆ No funding constraints were assumed.

# Assumptions

- **Custom Maintenance Assumptions.**

- ◆ This is maintenance of the non-COTS subsystems.
- ◆ For the new USGS system, assumes it is simpler than ECS, but still maintaining petabytes of data nearline.
- ◆ For supporting NASA's ECS, assumes:
  - An evolved ECS will exist.
  - NASA will maintain the evolved ECS.
  - ECS maintenance costs will also reduce 4% per year.
  - At least 2 other large sites will use ECS.
  - The USGS will fund it's share of ECS maintenance.
  - Numbers used are based on guesses, not competition-sensitive or confidential material.

# Assumptions

---

- **COTS Maintenance Assumptions.**
  - ◆ Based on a percentage of total hardware cost.
- **Reserve Assumptions.**
  - ◆ A 25% reserve is assumed.
- **Data Transfer Assumptions.**
  - ◆ The data transfer is assumed to happen via bulk transfer of tapes (the best possible assumption).
    - This has not been validated.
    - Provides significant risk to the New USGS System option.

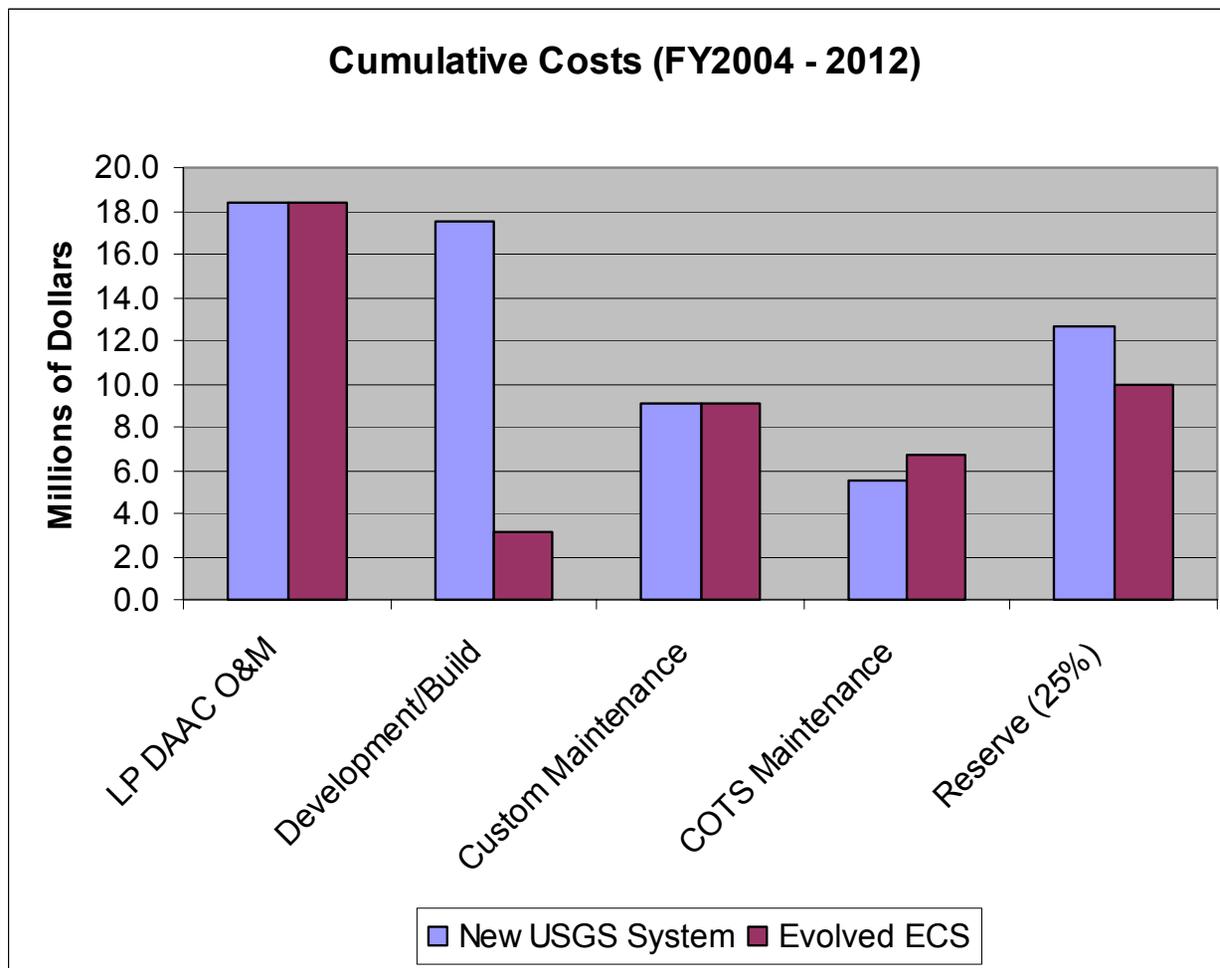
# Results - Estimated Costs

- **The purpose of this estimate was to address USGS costs, not NASA costs.**
- **Reasonable estimates and guesses were used, but not extensively validated.**
- **Accuracy target is plus or minus 50%.**
- **General idea was to determine:**
  - ◆ Whether the USGS funding required is around \$1M, \$10M, \$100M, or \$1B.
  - ◆ When the funding is needed.
  - ◆ What the major issues are.

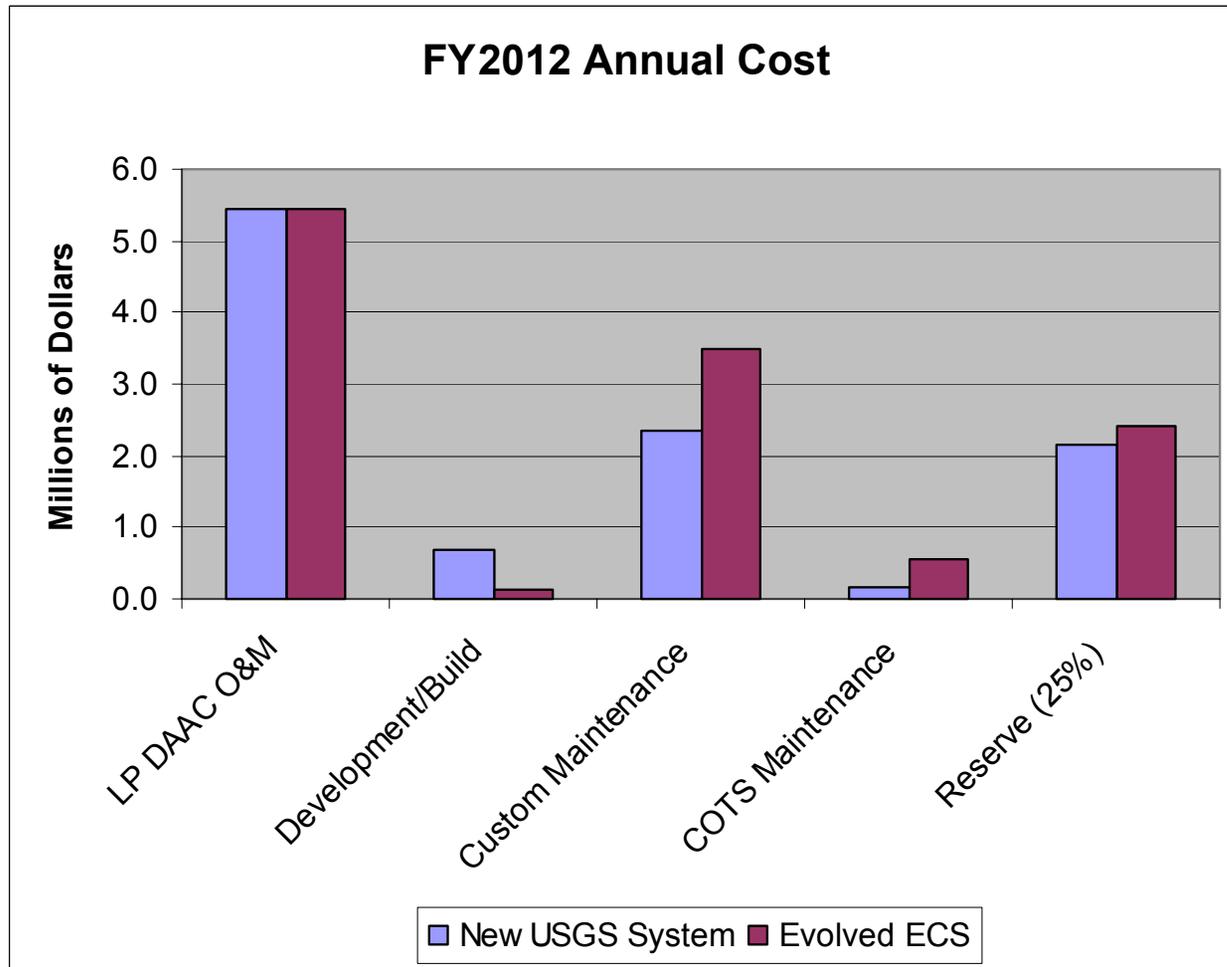
# Results – Estimated Costs

- **Between FY04 and FY12, the USGS cost is:**
  - ◆ \$63M for the new USGS system.
  - ◆ \$47M if sharing the evolved ECS system with NASA.
- **In FY12, the USGS annual cost is:**
  - ◆ \$11M per year for the new USGS system.
  - ◆ \$12M per year if sharing the evolved ECS.
- **The estimate's best case, with cheapest scenario, minus 50%, and with 25% reserve removed, is:**
  - ◆ \$19M for FY04 to FY12.
  - ◆ \$4.4M per year in FY12.
- ***This is certainly not enough, but still would be a challenge for the USGS to get appropriated.***

# USGS Cumulative Costs

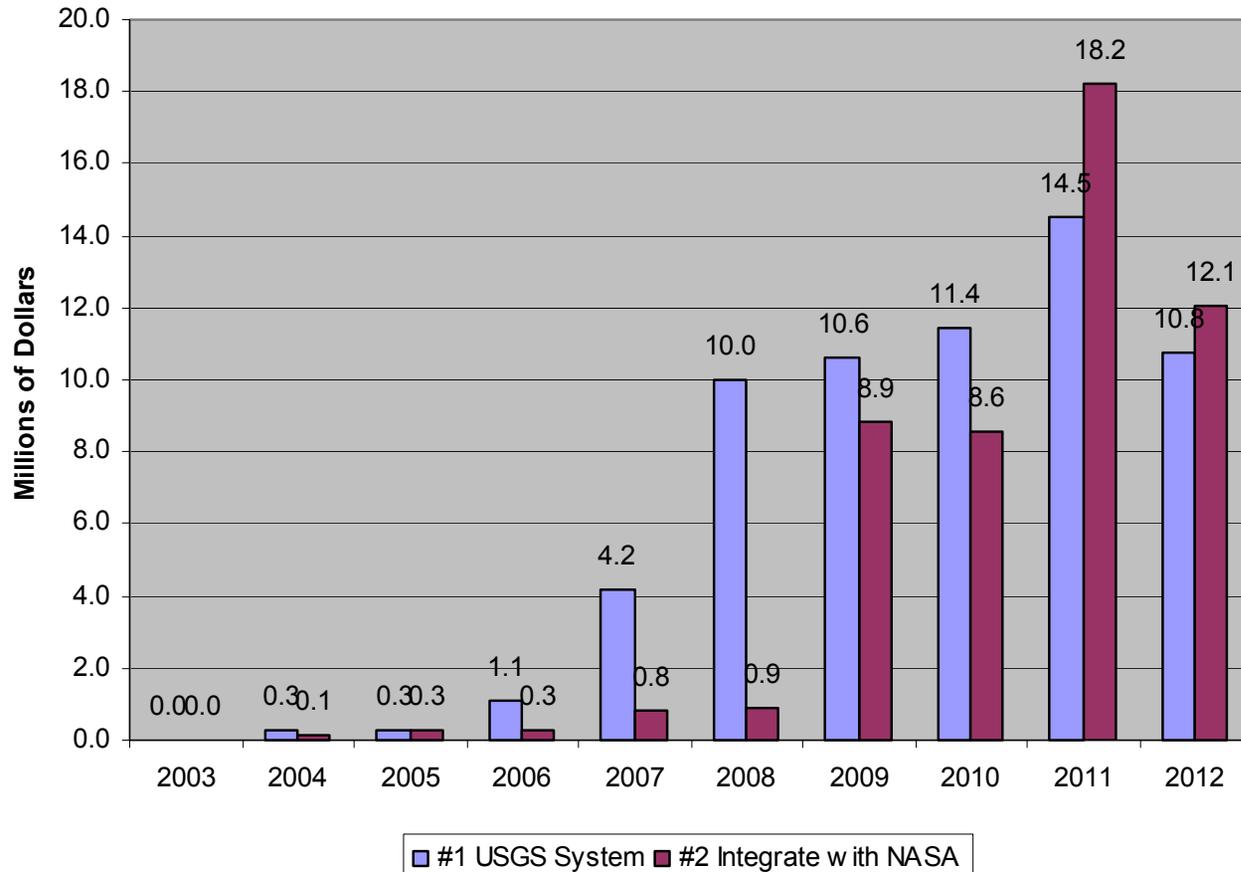


# Annual Costs



# Costs Per Year

Annual Funding Requirement for EOS Data (rough order of magnitude)



Land Processes DAAC  
Science Advisory Panel Meeting  
September 10 & 11, 2003

# Results – Major Points

---

- **It is approximately a third more costly to build a new USGS system compared to sharing an evolved ECS – before adding in any transfer costs.**
  - ◆ Most of the extra costs are development labor and new hardware purchases.
  - ◆ Given Moore's law, hardware costs are driven by schedule, so the early hardware purchase is expensive.

# Results – Major Points

- **However, the relatively low cost of sharing the evolved ECS strongly depends on NASA doing three things:**
  - ◆ Giving the USGS the ECS hardware for 3-5 years (at which time the USGS buys all new hardware).
  - ◆ Maintaining and using an evolved ECS at multiple sites.
    - If USGS fully inherited an evolved ECS in FY08, the cost from FY2004 – 2012 would be roughly a third more expensive than the new USGS system option.
    - Fully inheriting the ECS as it is today is unaffordable.
  - ◆ Maintaining and using an evolved ECS beyond FY2012.
    - Not impossible, given NASA philosophy with other major development projects, architectures, and infrastructure.

# Results – Major Points

---

- **The transfer schedule is very important.**
  - ◆ Funding takes years to acquire.
  - ◆ Systems take years to build.
  - ◆ Understanding and negotiating some participation in an evolved ECS would take at least a year.
  - ◆ Ends of mission affect reprocessing.
  - ◆ Reprocessing affects transfer schedules.
  - ◆ The estimated schedule is a good working schedule, but has not been validated.

# Results – Major Points

- **It is not clear how sensitive the costs are to the method, amount, and specific data sets transferred.**
  - ◆ Converting some data sets into process-on-demand would affect cost and feasibility.
    - Time passing favors processing on demand.
      - ◆ Hardware will be cheaper in the future.
      - ◆ Demand will decrease as the data ages.
    - Processing on demand can be funded by users in COFUR.
      - ◆ Alternatives are additional funding or not supporting product.
    - Avoids reprocessing issue and costs (for USGS and NASA).
    - Reduces archive and data management costs.
    - Adds software maintenance cost.
  - ◆ Reducing the number of bits transferred will not linearly reduce the funding required.

# Results – Major Points

- **Long-term archiving is different from short-term.**
  - ◆ User support goals are the same for both.
  - ◆ In LTA, ingest is an engineering job, not operations.
  - ◆ Reprocessing is generally not done by the LTA.
    - If data is still being reprocessed, it's not ready for the LTA.
    - LTA trades cost/benefit of new versions vs. new data sets - it would be very difficult for the LTA to justify new versions.
    - The LTA is willing to cooperate in reprocessing, though.
  - ◆ In general, higher level data is more perishable than lower level data, so it is less attractive to the LTA.
  - ◆ In general, the LTA likes to keep low level data and process higher level products on demand.
    - Keeping higher level products is done when processing on demand is impossible, costly, inconvenient to user, etc.

# How About Shared Data Sets?

- **The estimate so far was on separate data sets, which have one clear owner.**
- **Data sets with shared ownership, where perhaps NASA funds the three years since the last reprocessing, the USGS the rest, is possible.**
  - ◆ **Advantages:**
    - Work can be done in one facility that users can go to.
    - CDRs could be built to use 30+ years of data.
    - It would foster closer cooperation between the two agencies.
    - There is a potential for the shared facility being more responsive to the wider land science community.
  - ◆ **Disadvantages**
    - The devil is in the details: who owns, controls, and funds the systems and the data.
    - CDR reprocessing schedule is very much TBD.

# Where Do We Go From Here?

- **The transfer schedule is very important.**
  - ◆ The USGS requires years of lead time.
    - Some work will be required regardless of scenario:
      - ◆ Science justification and community involvement.
      - ◆ Funding work in USGS, DOI, and Congress.
    - System work may take years as well.
    - A “need date” is required for most of that work.
  - ◆ However, NASA often cannot accurately predict the end of missions.
  - ◆ Reprocessing is also TBD and drives schedule.
  - ◆ The USGS does not want to (and probably cannot) go through the significant effort required to acquire new funds without good schedule information.

# Where Do We Go From Here?

- **What do we transfer?**
  - ◆ Science and programmatic justification is needed in FY04.
- **When do we transfer it?**
  - ◆ The date of the last processing or reprocessing is needed.
    - Today this depends strongly on the end of mission.
  - ◆ Worse case is to wait until the last reprocessing, then start pursuing funding (could take 4 to 6 years from that point).
- **How do we transfer it?**
  - ◆ What and when must be known or closely estimated.
  - ◆ Use of evolved ECS depends on cost savings vs. risk of the USGS fully inheriting the system.
  - ◆ Discussion of CDRs and a possible closer relationship between NASA and the USGS should be pursued.
- **Who and where are already agreed to.**

# What Do We Do Now?

- **Both agencies should / might consider:**
  - ◆ Immediately work “what, when, and how” on their own.
    - See Part 2 for details on “what and how”.
  - ◆ Start discussing the subject together.
    - As agreed to in the original MOU.
    - Particularly if the shared data set ownership model is desired.
- **USGS should:**
  - ◆ Work hard to refine / lower estimate as much as possible.
  - ◆ Prepare for a major funding initiative in FY04.
- **NASA might consider:**
  - ◆ Participating in the USGS funding initiative.
  - ◆ Factoring in that ECS evolution, CDRs, and need dates will affect LTA.
    - The more uncertainty there is in all of these, the more likely it is the USGS will pursue a new USGS system.

# What Do We Do Now?

---

**The Science Advisory Panel should...?**