National Aeronautics and Space Administration

White Paper: Resident Archives in the S3C Data Environment

Prepared by Ed Grayzeck and Don Sawyer (NSSDC)

July 8, 2005 update

1. Preface
2. Introduction
3. Definition of a Resident Archive
4. Implementation
5. Role of NSSDC
6. Summary
7. Figures

8. Appendices

1. Preface

This white paper was solicited to provide context for a discussion to establish the Resident Archive (RA) process and determine the role of NSSDC. It draws upon SEC's history and experience, the strategic and implementation plans of NASA and its organizations, trends in the development of data systems, technology, and standards. These ideas were presented by Chuck Holmes in his report to the S3CDCWG meeting in April 2004. In figure 1, he outlined the current state of the data life cycle. He also sketched an RA started via an "archive arrangement" managed by NSSDC (figure 2).

To flesh out the initial outline, a series of open meetings were held with the Living With a Star Data Environment Working Group (LWS DEWG, e.g., A. Roberts, A. Szabo, J. Gurman), the SEC Active Archive (SECAA, e.g., R. McGuire, R. Candey, J. King, S. Fung), HQ sponsors (e.g., C. Holmes, J. Bredekamp), and other invited scientists. From these three sessions plus the email correspondence, a strawman scenario was assembled and will be described later. There was some concern that the missions already are providing data access and should be doing archive preparation for mission end, so it has been questioned why is an RA needed. The RA reflects the reality of how SEC missions work. By extending the mission beyond phase E(or F) for an RA, the community gets the best quality final data (perhaps as new products) which are served by the mission experts (who provide a unique functionality). The general consensus was that an "archive arrangement" have both a section on implementation and termination of the RA. The group further recommended that any data plan should be composed early by the missions, and they address the need to transfer the data of primary science potential to a long-term archive so that the data can be accessed and managed in a cost effective manner. (Chuck indicated that an earlier ruling deemed the PDMP as a project document so an "archive arrangement" would be separate from the PDMP, and modified till the RA is established.) The point was emphasized that we need to get the missions to seriously ensure archiving and making the data usefully available, e.g., through Virtual Observatories (VOs). It was judged important to get the PIs to fulfill their responsibilities.

2. Introduction

In the case of NASA space science missions, the National Space Sciences Data Center (NSSDC) at NASA Goddard Space Flight Center is the permanent archive for that data.

US missions in astrophysics, often work through existing data centers to host the data archiving and distribution. Such is the case for Galaxy Evolution Explorer (GALEX) which has its Science Operations Center (SOC) at the California Institute of Technology but distributes its data through Multi-mission Archive at Space Telescope Science Institute (MAST). In this approach, the mission creates a remote archive to make use of the available tools for viewing the data and answering queries. In a different approach, the recent ESA Infrared Space Observatory (ISO) mission is currently in an "archive phase" during which there will be an effort to consolidate the knowledge gained through the mission lifetime by co-locating the ISO Data Center at the operations center which is now the European Space Astronomy Center (ESAC).

In the solar system missions sponsored by NASA, the data must be archived through the Planetary Data System. In a "PDS supported" mission, an archive scientist from one of the lead scientific nodes (called a Discipline Node) works with the mission to get the data in archive form. In certain cases the PDS works directly with the missions to establish a "data node", wherein the data is immediately available to the community through the mission led institution. The data node can be supported by mission funds or PDS funds; in either case, the data node reports to the appropriate Discipline Node with which it has signed a Memorandum of Understanding (MOU). At the close of the mission (or upon other arrangements), the data node ceases to exist and the data is transferred to PDS seamlessly. Normally, the planetary missions are PI led with small Co-I or instrument teams. The Planetary Data System, in their Life Cycle approach, now calls for the Archive Plan to be included in the PDMP.

For research areas involving the Sun Solar System Connection (S3C), loose collaborations by missions are typical. To study the sun, the SOHO mission has proven successful with its shared approach between the European and US space agencies for both operations and archiving. Other collaborative space physics mission such as Ulysses and Cluster have similar methods for making the data available. In the S3C community, the mission lifetime is sometimes extended through a post operations phase which was proposed for the SAMPEX mission. During the proposed "SAMPEX Data Center" period, the data is prepared for archiving. SAMPEX is a recent example of how the Resident Archive process might operate; other examples include the Yohkoh archive maintained by Montana State University which is associated with the Virtual Solar Observatory (VSO) at GSFC.

Usually, the scientific data are prepared by a small group within the experimenter's team supported by 'archive scientists' from an archive/science specialist group which is an active archive or NSSDC.

The experimenter team lead by the Principal Investigator and consisting of the staff of his/her institute and sometimes a large number of Co-Investigators, wants and needs to look at the acquired data as soon as possible after it's availability on the ground. These early investigations are often an important part of the instrument operation and the data validation. As the data archive pipeline is usually not ready – and not intended to be ready – at this time, the experiment team bases it's analysis on data in telemetry or some intermediate file format, often called working archive. Requirements from the experimenter team on processing, visualizing, searching, browsing and calibration support are very often not linked to the archiving effort and solved by implementations that are only available to this experimenter team and herewith lost as soon as the teams loose interest or run out of mission funding. Any other interested scientist depends on the good will of the Principal Investigator to use these tools and services – if they are still available.

NASA requires that a mission's data management plan be a total package end-to-end, and that serving of data and archive planning be addressed early in the project. In most cases, SEC missions do not address their archiving responsibilities in the operations phase (E). In the RA methodology, there would be an "archive arrangement" (called a Resident Archive Data Plan - RADP, an update to the PDMP) which is proposed as the mission evolves and is tied into the senior review cycle. The RADP will take effect if the mission is terminated.

3. Defining a Resident Archive

During the discussions with the LWS DEWG and SECAA, six prioritized functions were identified as follows:

1. Produce as complete a set of data products as possible (either new or improved, comprehensive, high time resolution, high quality) to the stage were they can be served;
2. Ensure that the mission data are served to the general space and solar physics community in an efficient and scientifically useful interoperable manner consistent with community data environment standards (e.g., VOs) and using readily sustainable, automated software;
3. Maintain the integrity of the data by safeguarding against data loss which could be effected by providing a mirror site, e.g., NSSDC;
4. Document the above (including mission and PI information) as required to maintain independent usability;
5. Obtain community feedback on the above to insure success;
6. Make sure that the data will be archived after the RA is no longer needed (e.g., preserved by transferring to another RA, or NSSDC).

The RA is expected to have two modes of operation; a startup phase and an operational phase. In the former phase, the above six functions are the prioritized set, while in the latter phase, the development of new/improved products becomes secondary to the continual serving of the data so that the community finds it useful. There will be reviews for these stages and a strawman list of relevant criteria were developed as follows (note some criteria still have open questions that need addressing):

1. Bring a complete set of new/improved data products that are served to the community

data products are frequently downloaded
data products are supported by the community as essential.
data products are served using readily sustainable, automated software
(automated software needs definition or examples from the community)

2. Provide data products that are usefully available

data are not in proprietary format
data are provided continuously by the resident archives
data served to the community in a useful manner
data in formats preferred by the community
(who chooses the preferred format set SECDCWG or a community effort)

3. Maintain the integrity of the data as a resident archive

provide data backup if needed or off-site storage
provide for a mirroring the data, e.g., to NSSDC
transition plan in place for catastrophic problems
(what is a reasonable cost for data security?)

4. Provide documentation to maintain independent usability

description of mission, spacecraft including history
description of instrument and/or user's manual
data information including a) processing history (levels),
b) coordinate systems, c) parameters, d) limitations
calibration techniques and associated files
ancillary files that may be needed for interpretation
tracking and ephemeris information
(should software be archived and at what level of support)

5. Interact with the community regarding data quality and services

journal citations
data usage (file transfers, executions)
log complaints and eventual resolutions
look at data availability as described in the RADP
(who develops metrics to reflect science community both in breadth and usage)

6. Termination plan

(At termination the data generally will continue to be automatically served)
the delivery mechanism is identified for the final disposition of the data
the responsibilities for the delivery and acceptance are put forward.
data content, formats, and volume are clearly specified.
(what is a realistic timeline for final archiving)

Finally, a set of derived requirements were drawn as follows:

Provide procedures so there is no scientific data content lost during RA phase
Provide expert knowledge to best use the data, for continuous serving to the community
Provide for improved data quality which will be preserved long-term

4. Implementation

Resident Archive User Group

It is expected that the RA status will be populated by a number of projects, initially starting with SAMPEX (ACE) and Yokhok. A Resident Archive User Group (RAUG) should be formed independent of NSSDC. The purpose of the RAUG includes refining the criteria used in the peer review process, sharing common problems and solutions such as the method to develop metrics for usage of data, evolution of services such as virtual data products, and the priority for RA distribution of funds. In the strawman model (figure 3), NSSDC organizes the RAUG but remains an ex-officio member. Furthermore, the NSSDC manages the HQ grants to the individual RAs with the lines of communication established as shown in that figure. The RAUG membership will be open to all SEC missions; a newly formed RA will be expected to be a member. Details of the RAUG composition, term and mechanics are outlined in Appendix A.

The startup mode of an RA.

A typical Senior Review (SR) is called by HQ with missions (PIs) expected to provide an RADP as part of the mission proposal (see figure 4). The RADP shall include an RA termination plan for archiving, e.g., with NSSDC. If a mission is given a low priority, and it is slated to cease operations (typically one year), then the RADP may go through iterations with HQ and possibly with NSSDC. The time frame is to be no more than 3 months duration, so that the Resident Archive proposal can get funded without interruption. A startup review is called by HQ to judge the proposal (NSSDC only advises) to determine if the group is ready to take on the functions of an RA and that the scope of data products is both timely and feasible. Part of this process is for the group to develop an implementation strategy that basically is the "How to " manual for their specific RA (Appendix D) which lays out methods to meet the six functions. In the startup mode the RA has both a scientific and technical role. It is expected that startup costs for an RA will depend on the existing infrastructure the range of new/improved data products that are proposed but will be a few FTEs/year.

If the mission, via the PI, chooses to not pursue the RA phase, they must notify NSSDC within 3 months of the SR published results. (The serving of data to the community must continue uninterrupted.) If the mission does not opt to become an RA, HQ shall initiate (with NSSDC input) to solicit other Co-Is of the mission to propose for the RA status; only as an exception will an outside (the mission) group be solicited (such as an existing RA) to seek RA status to curate the respective data. Such an entity, the mission, NSSDC, and HQ will collectively discuss a new RADP. Once an RADP has been accepted then it is subject to a review (HQ calls the review with NSSDC support), and if the RA is approved, funding will be issued. The mechanics of running the startup review are outlined in Appendix B.

The Operational Mode of an RA

Once the RA is selected, quarterly reports are expected along with participation in the Resident Archive User Group (Appendix A). The performance of an RA will be judged through a peer review process that will consider the six qualities listed before that are used to judge the performance of the RA. HQ funds the RAs but NSSDC will serve as the manager to negotiate and monitor grants which is assumed to start at the beginning of the RA (figure 2). The RA can update the RADP during its lifetime in consultation with NSSDC . It is expected that a schedule for data delivery will be set, preference being for periodic transfers. The mechanics of running a performance peer review are given in Appendix C.

Since in the operational phase, the new/improved products will have been completed, the RA mainly serves a technical role and the criteria are reprioritized accordingly. It is expected that the operational phase for an RA will need less than 1 FTE/year. Normally, the RA terminates when the data usage has diminished or its expert personnel is no longer sustainable. At that point there are two options. The data can go to a permanent archive such as NSSDC, or it can be transferred to another RA or similar Consortium of RAs (CRA) that have been formed to focus on an instrument or discipline. The goal is for any transfer to be seamless. If only RA web services are considered, mirror sites (NSSDC, another RA, CRA, or commercial) make the change transparent.

As part of the community discussions, we brought up the concept of how the RAs may evolve, namely, the SAMPEX Data Center (now at the University of Maryland) will become part of the ACE Science Center; there is a coalescence which is largely driven by expertise and economy of scale. The natural co-location of RAs would be an example of how a Consortium of RAs (CRA) might form. We decided that the grouping of RAs by discipline or instrument was a good thing as long as such grouping provided a unique function and was cost effective. If there were to be a "performance peer review" for the SAMPEX RA, it could be judged separately from a possible ACE RA, since the former data products are stable. In the case of Polar (distributed in figure 3), an RA organized around instruments seemed reasonable. An RA fits the VO concept since it continues to serve the data once the mission terminates according to community standards. It is expected that there could be a few CRAs but the number of CRAs and RAs ought to be limited, so that they are cost effective.

5. Role of NSSDC

What should be the role of NSSDC? The NSSDC should take the lead in starting a process to gather community input on preferred standards. This could include recommendations on data formats. It could/should involve the fundamental question on what does it take to make the data independently useful.

The NSSDC has been working closely with the Planetary Data System on an efficient electronic method to deliver data and can leverage that experience to make any data transfers a seamless one to the community.

NSSDC has worked with many missions in the preparing and reviewing of PDMPs and a Guideline for PDMPs is available under Archive Support at its web page. The NSSDC could expand these guidelines to formulate an RADP example based on a similar approach used by PDS, namely, an Archive Generation, Validation, and Transfer Plan.

The NSSDC roles needed to make the RA process work are as follows:

Implement the RA concept for HQ
Assist the missions with devising an RADP
Review the RADP for individual missions
Manage the individual “grants” for the RAs and/or CRAs
Assist with the solicitation for an RA when requested by HQ
Organize and support the RAUG and provide for meetings
Assist with the termination plan for an individual RADP
Provide recommendation from the startup and performance reviews

NSSDC current management can step up to many of these roles. In addition, it has recent experience from the PDS on how to organize peer data reviews (using the NSSDC based Electronic Handbooks), manage data nodes through grants, and form advisory councils (similar to user groups) for the nodes.

6.Summary

The general consensus was that the Resident Archive dealt with a real problem in SEC missions but that it was prudent to have the discussions about the RADP early in the mission cycle (during assembly of the PDMP as shown in figure 4) so that data could be served in a most useful and effective manner before the RA stage. To be a cost effective approach to archiving data, the number of RAs (CRAs) needs to be limited since it is estimated that the cost of each RA (startup and operational) can be significant. To get a handle on costs, we site a similar example.

In Earth Science there has been a Strategic Evolution of ESE Data Systems (SEEDS) study to characterize the 20 operational Data Centers to plan for startup and operational costs for any PI led Data Service Provider (DSP). The resultant cost estimation (CE) tool is based on a comparables database which shows a wide variation depending on Level of Service (LOS) but the bottom line is that even for the simplest case of a store and forward one (Remote Sensing Systems), the effort is still about 0.9 FTE/DAAC/year.

One outcome could be that NSSDC take on a pilot project to assist HQ in the management of SAMPEX (or VSO for Yohkoh) to get realistic costs for both the startup and operational phase, and then implement this process, if feasible, in 2005.

Figures

Current model for data life cycle

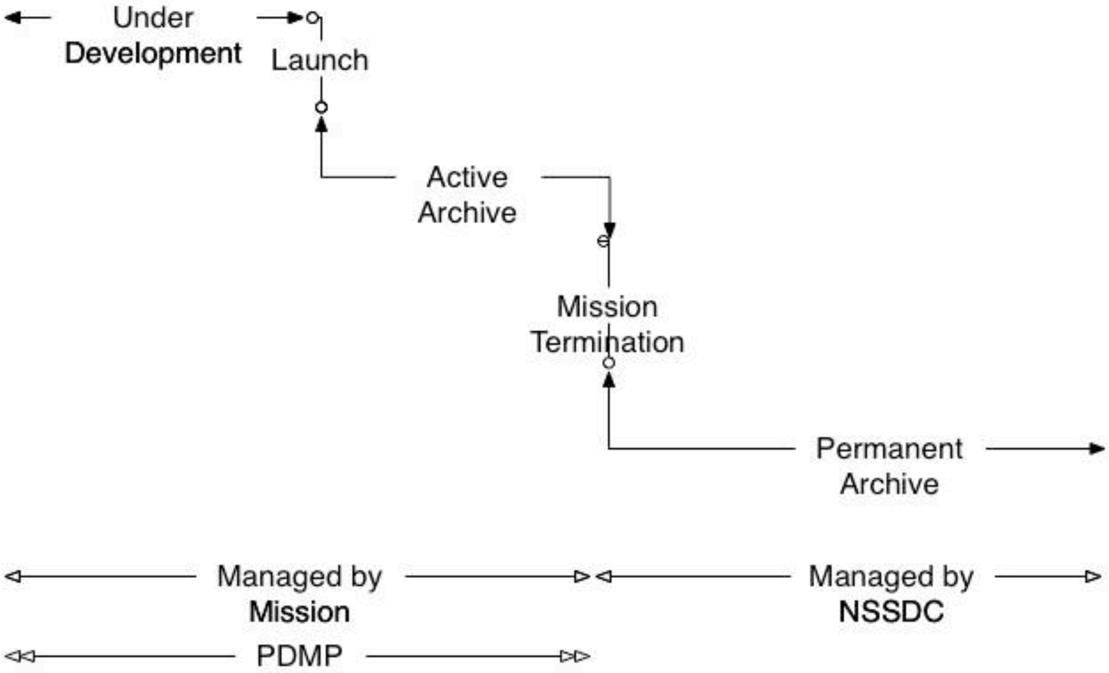
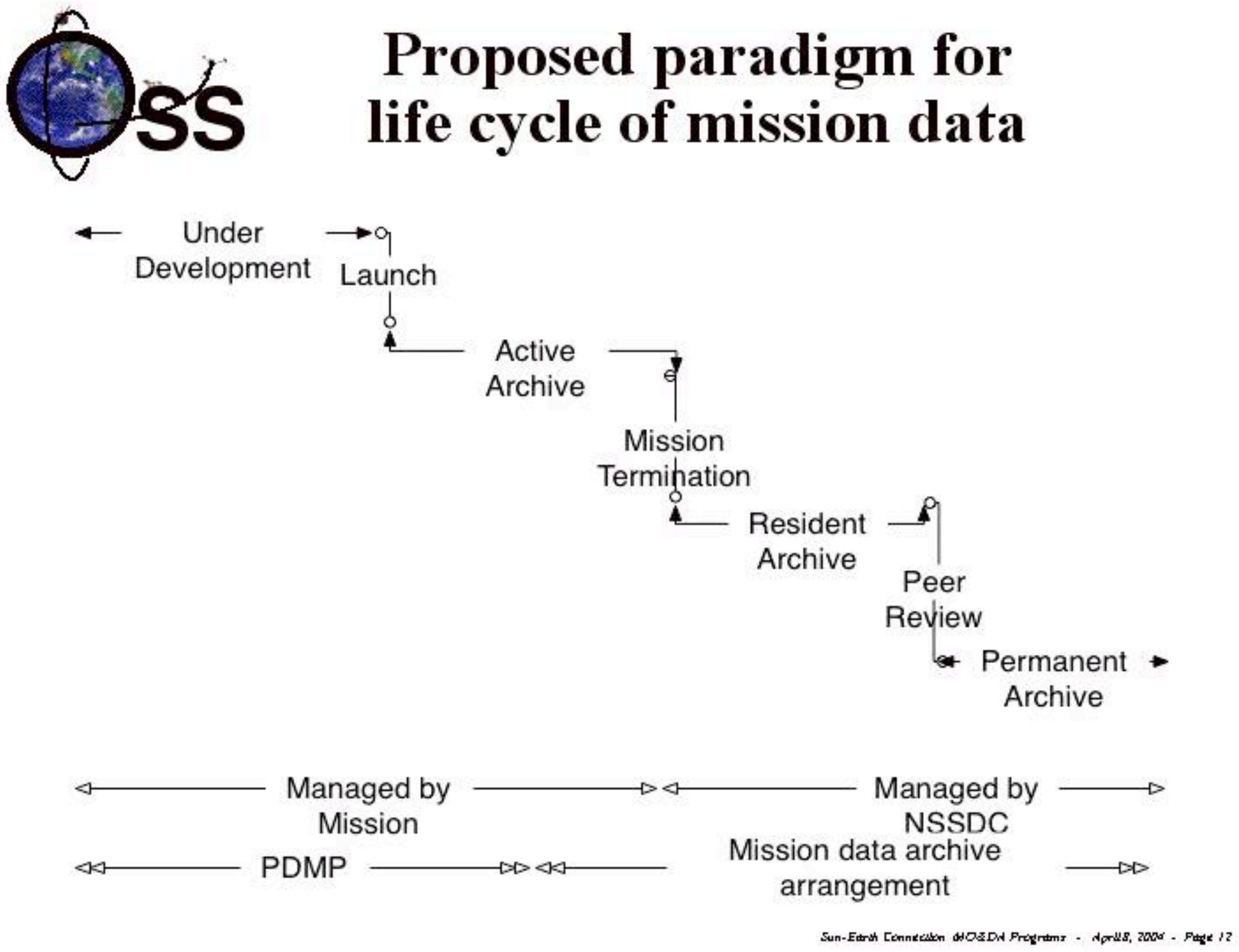


Figure 1: Current SEC model of data life cycle (C. Holmes, Apr 2004)



Proposed paradigm for life cycle of mission data

Under Development

Launch

Active Archive

Mission Termination

Resident Archive

Peer Review

Permanent Archive

Managed by Mission

Managed by NSSDC

PDMP

Mission data archive arrangement

Sun-Earth Connection MO&DA Programs - April 8, 2004 - Page 12

Figure 2: Proposed paradigm for life cycle of mission data (C. Holmes, Apr 2004)

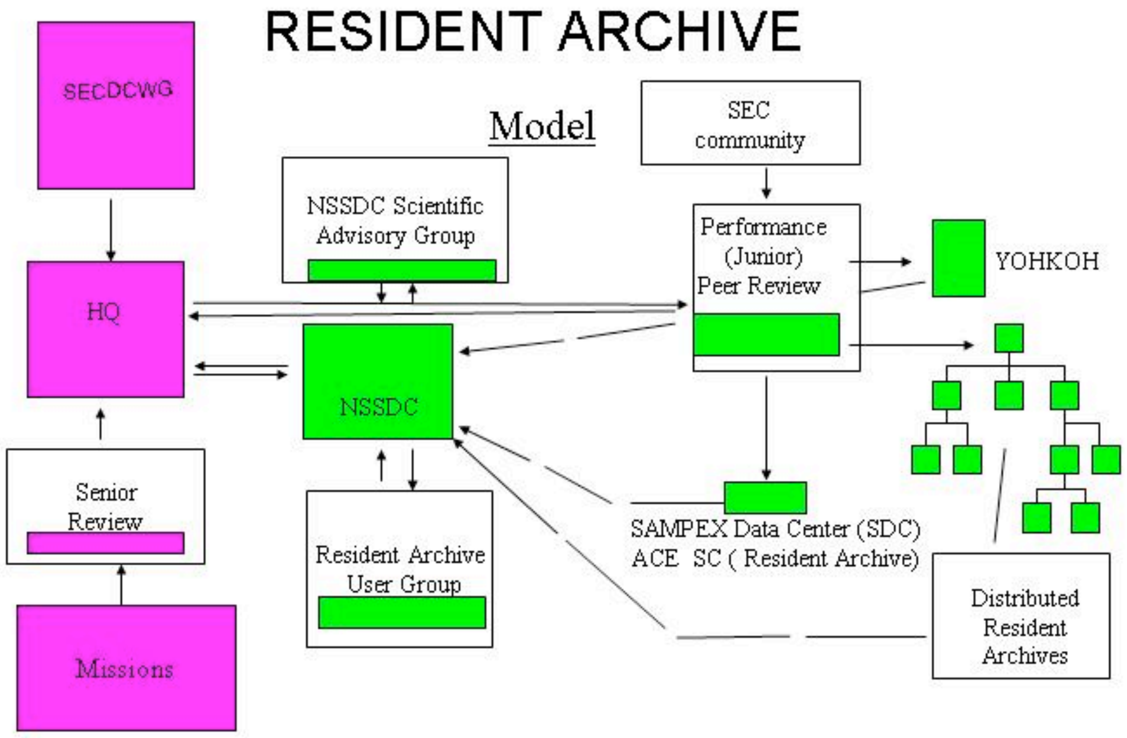
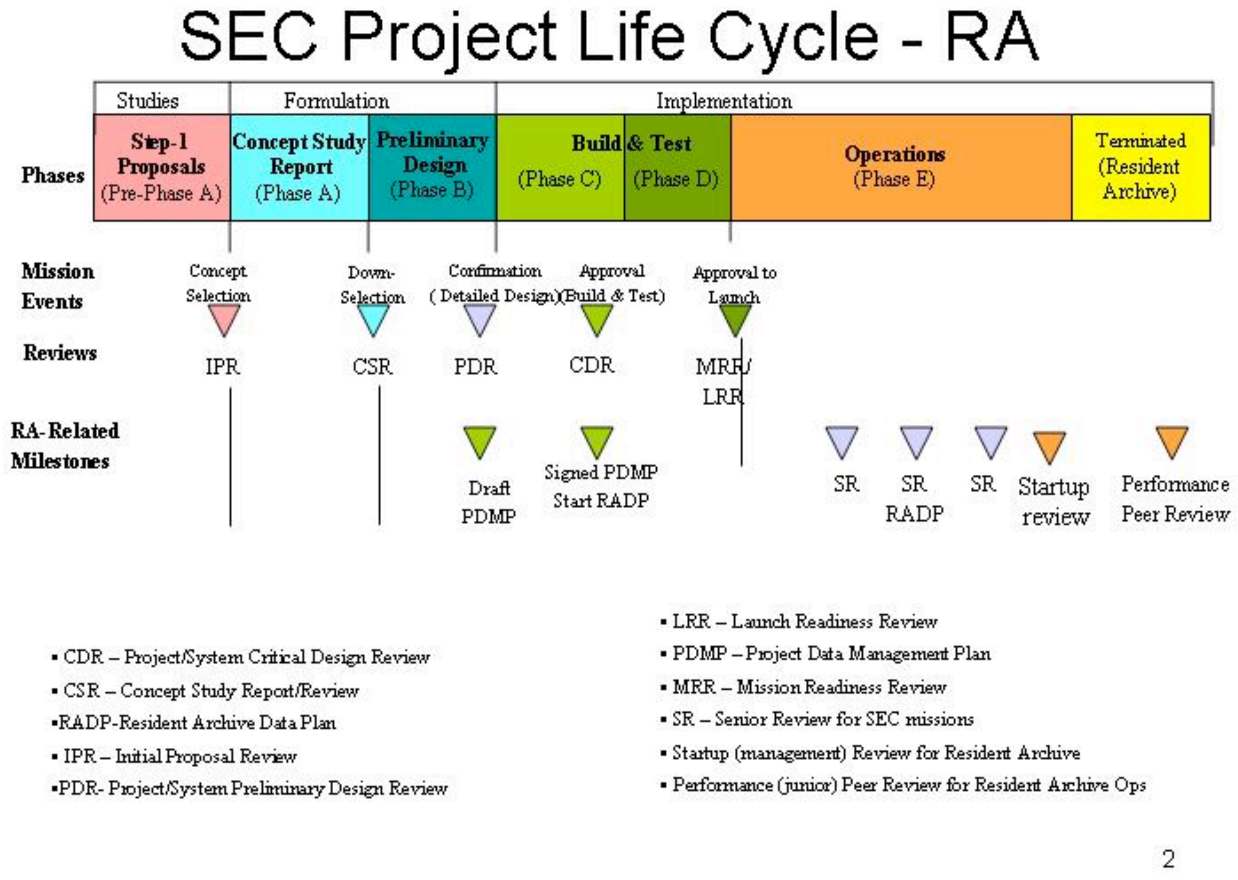

RESIDENT ARCHIVE
Model
SECDCWG
HQ
Senior Review
Missions
NSSDC Scientific Advisory Group
NSSDC
Resident Archive User Group
SEC community
Performance (Junior) Peer Review
YOHKOH
SAMPEX Data Center (SDC)
ACE SC (Resident Archive)
Distributed Resident Archives

Figure 3: Strawman model for Resident Archive and the role of NSSDC



SEC Project Life Cycle - RA

Studies
Formulation
Implementation

Phases
Step-1 Proposals (Pre-Phase A)
Concept Study Report (Phase A)
Preliminary Design (Phase B)
Build & Test (Phase C) (Phase D)
Operations (Phase E)
Terminated (Resident Archive)

Mission Events
Concept Selection
Down-Selection
Confirmation (Detailed Design)
Approval (Build & Test)
Approval to Launch

Reviews
IPR
CSR
PDR
CDR
MRR/ LRR

RA-Related Milestones
Draft PDMP
Signed PDMP Start RADP
SR
SR RADP
SR
Startup review
Performance Peer Review

- CDR – Project/System Critical Design Review
- CSR – Concept Study Report/Review
- RADP-Resident Archive Data Plan
- IPR – Initial Proposal Review
- PDR- Project/System Preliminary Design Review

- LRR – Launch Readiness Review
- PDMP – Project Data Management Plan
- MRR – Mission Readiness Review
- SR – Senior Review for SEC missions
- Startup (management) Review for Resident Archive
- Performance (junior) Peer Review for Resident Archive Ops

2

Figure 4: SEC project life cycle showing Resident Archive milestones

Appendices

Appendix A

Resident Archive User Group (RAUG) similar to a science working team

Possible composition: Missions, RAs, CRAs, NSSDC as advisor

Purpose: refines the criteria for evaluating the RA functions (Web/ftp) ?
shares common problems/solutions such as metrics to report usage
identifies the evolution of services such as virtual data products
advises on the priority for distribution of funds

structure: informal group when RAs few in number
invite all SEC missions to participate
meets annually, usually at scientific meeting
goals: work with NSSDC to provide RA status
cooperate to develop future RAs, CRAs

Appendix B

Mechanics of Management Startup Review

“startup review” composition: 3 members – HQ, non-RAUG (NSSDC only advises)
"startup review" judges on the six core functions of an RA, and the RADP
meet 1/2 day for individual RA proposal
when more RAs exist, can consider them as a group for prioritization

Appendix C

Mechanics of Performance Peer Review

performance peer review composition: reviewers from the community, SECDCWG, SR?,
with SECAA and NSSDC support
grouping of RAs (CRA), RA judged separately on 6 functions (modified prioritization)
meets 6 months before scheduled SR or at least bi-annually
report to HQ, NSSDC submits separate report for concurrence

Appendix D

Implementation strategy for a Resident Archive associated with NSSDC

1. Purpose of the <mission> <data set> Resident Archive
   a. The NSSDC manages a Resident Archive for a limited time period to prepare data to distribute as well as to prepare the data sets for archiving. The HQ and/or NSSDC can accept a proposal for the Resident Archive phase any time during a mission's lifetime but will actively solicit a proposal during the termination year of a mission. NSSDC expects a Resident Archive to exist for at least the duration of the proposed Resident Archive Data Plan in which the data are finalized (or longer, based on funds)

   b. A Resident Archive reports to and receives assistance from the NSSDC; funding is through HQ.

   c. The Resident Archive undergoes periodic reviews to ascertain if the updated data products are acceptable to the community, if the services are adequate, and if the value-time profile for the data justifies continued funding of the Resident Archive.

   d. By the end of the time period the Resident Archive must provide for final archiving of the data sets as specified in the proposed RADP.

   e. All community archive standards and processes apply to the Resident Archive data

2. Services Provided By the Resident Archive

   a. Provide (and track) web access of the data otherwise done by the Mission

   b. Prepare data (transform if needed) to community accepted formats

   c. Provide additional services that add value to the data set through mission expertise that includes documentation so the data is independently useable

   d. Interact with community data requests/comments on data products and services

3. Other Responsibilities of Resident Archive

   a. Quarterly reports to NSSDC that show adherence to archive standards

   b. Participate in bi-annual "performance peer review" administered by NSSDC

   c. Plan and implement transition of data and services at end of performance period

4. Responsibilities of NSSDC

   a. Develop community standards for archiving, formats, compliance, etc.

   b. Validate data and documentation delivered by the Resident Archive

   c. Help RA with transition at end of performance period

5. NSSDC oversight elements

   a. quarterly reports should also focus on requests and level of service provided

   b. data usage as recorded by file transfer or executions logged to IP addresses

   c. data products developed as in the RADP that are available and useful

   d. integrity for the data site needs to be provided, e.g., backup or mirror sites