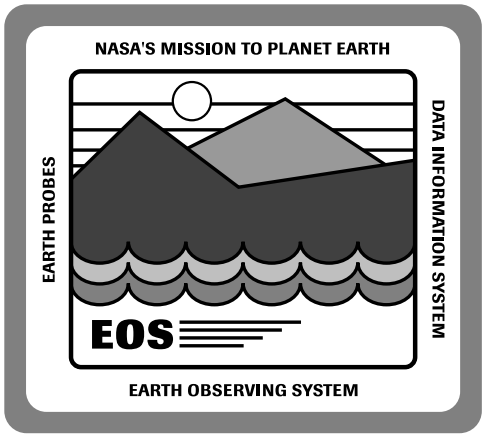
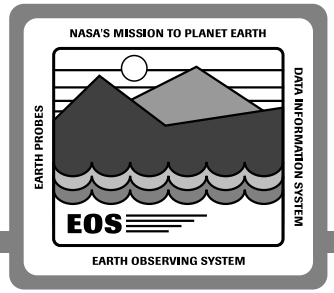


Understanding Science User Concerns

Bob Curran

13 - 14 December 1993

Understand Science User Concerns



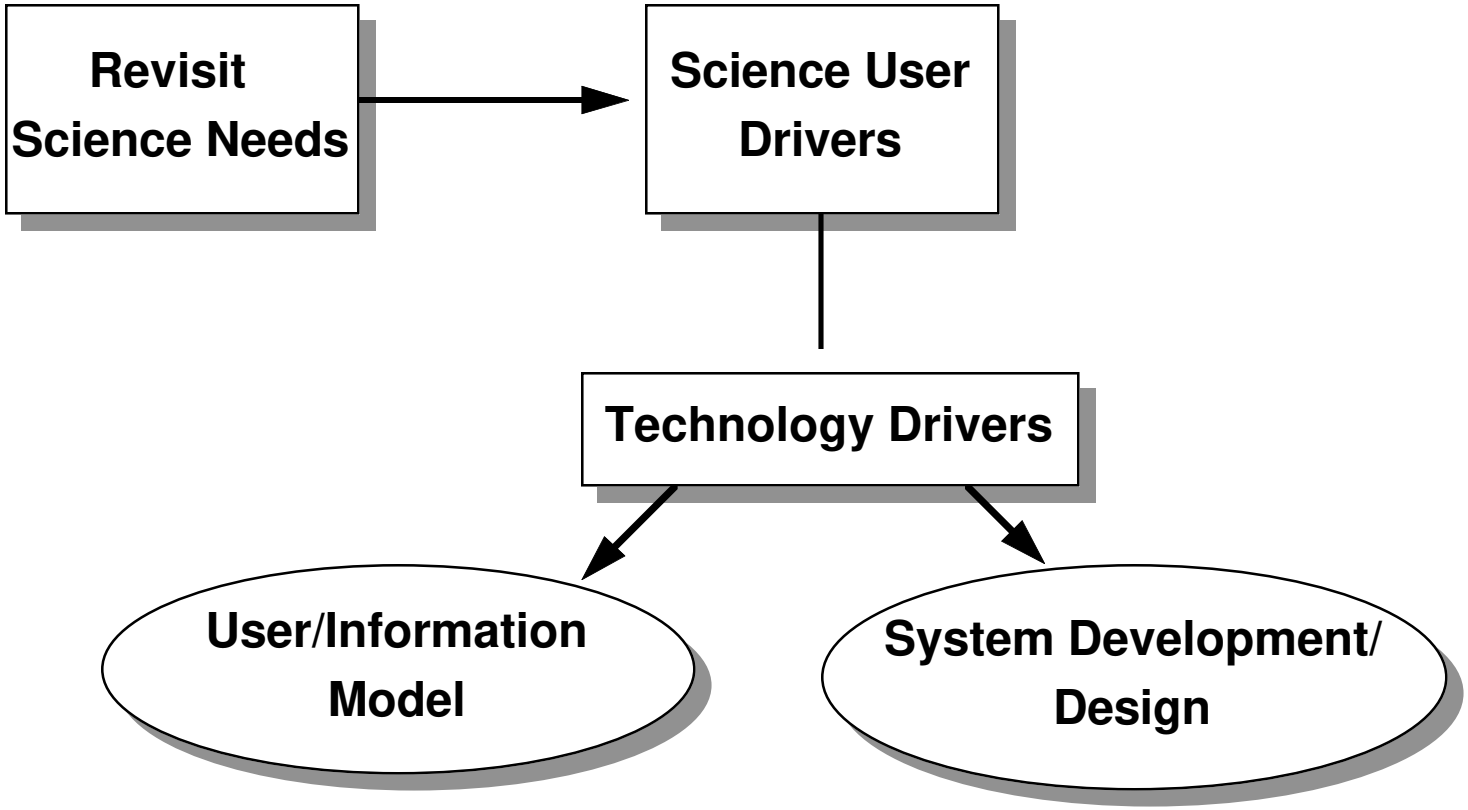
Discussions with Science Community

Objective

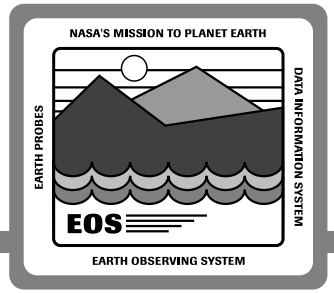
- **Better understand scientists concerns with the development of ECS**
- **Understand the Earth science research driven basis for data and information system needs**
- **Identify areas for future collaboration in "problem solving" and prototyping**
- **Ensure that scientists data and information system needs are accurately translated into analyses and design approaches which impact system development**

ECS visit team - Earth scientist, computer scientist, and where possible system engineer and or development segment representative

From Science User Needs to Information System Development



Schedule of ECS Visits to Science Investigation Institutions



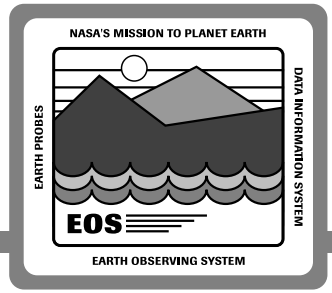
Data panel members

9/29/93	Bill Emery	Colorado
9/29/93	Paul Rotar	NCAR
9/30/93	Roger Barry	Univ. of Colo.
10/21/93	Jeff Dozier	UCSB
11/23/93	Bob Evans	Univ. Miami
12/07/93	Dave Glover	WHOI

Interdisciplinary Science Investigators

9/24/93	Bruce Wielicki	LaRC
10/12/93	Berrien Moore	UNH
10/15/93	Mark Abbott	OSU
10/18/93	Ricky Rood	GSFC
12/03/93	Dickinson/ Sarooshian	Arizona

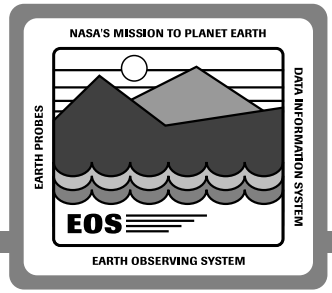
Science Drivers



Science (user) drivers: Issues associated with scientist's direct interaction with the system, or with concerns about how the system might limit or promote future scientific analysis.

- Support a dynamic product life cycle and easily extensible product set
- Support a fully interactive investigation capability
- Support user-to-user collaboration
- Facilitate an efficient search and "access" paradigm
- Support an information-rich data pyramid
- Support the integration of independent investigator tools
- Provide distributed administration and control

Science Drivers



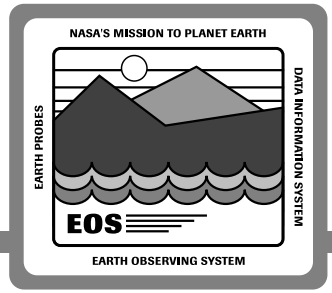
Support a dynamic product life cycle and easily extensible product set

- **Data "products" are in actuality a mixture of analysis and modeling (pixel "mixing" models -- e.g. atmospheric correction models, dispersion models, etc.). Scientists are continually refining these models to provide more accurate products. These products will vie for "shelf space" in the ECS data "supermarket", with newer products replacing older ones on an ongoing basis.**

Support a fully interactive investigation capability

- **This driver reflects the scientists' desires for a streamlined environment in which data is readily available for interactive investigations. "Interactions" can include combinations of simple display of images, more complex data manipulations and visualization, and algorithmic processing (e.g. to attempt to extract new features)**

Support a dynamic product life cycle and easily extensible product set (Continued)



(Emery - Colo.) **... the predominant requests in the scientific research community will be for level 1B data, not higher level products. Researchers should be able to provide ("publish") new products to the system. New products that generate a lot of interest should then be "migrated" to the DAACs for routine production.**

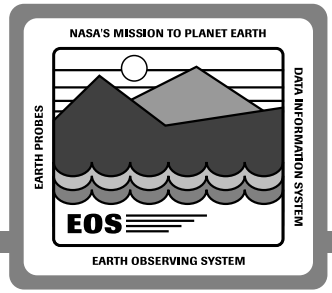
(Moore - UNH) **...sometimes the SCFs might be the best place to produce data products since they have more computing power than many DAACs and they are the source of the algorithms that transform sensor data into products.**

... data product creation might move over the life of ECS , from the SCF to the DAAC, back to the SCF when an algorithm is fixed or improved, and then back to the DAAC.

(Abbott - OSU) **Abbott subscribes to Stonebraker's philosophy of "all you need is the recipe."**

(Richman - OSU) **There is often a lack of unanimity in the science community about the data processing algorithm or the application of that algorithm to a specific problem. It is unacceptable for the processing center to offer the product to the researcher in the processed form (after application of a particular algorithm selected by the ECS).**

Support a dynamic product life cycle and easily extensible product set (Continued)



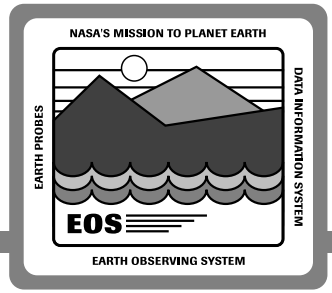
(Abbott - OSU) For his work, Abbott stresses the need for the availability of raw (unprocessed) data because the generation of quality high level products changes dramatically with time. A scientist will want to apply his own tools, but when he uses complementary data (e.g. ozone field distributions) he'll want to be able to pull in the "latest and greatest". This means communicating with a high power, yet flexible system.

(Abbott - OSU) There should be a few core products that one can always find in a predictable location, and other than that, products vie for shelf space depending on some form of "economic value". Abbott's idea of economic value is some measure of use by the scientific community.

(Glover - WHOI) Does not want to be data producer because of administrative responsibilities

(Evans - U. Miami) Because of the complexity of product generation cycle (ocean color) some products need "hands-on" processing at the SCF

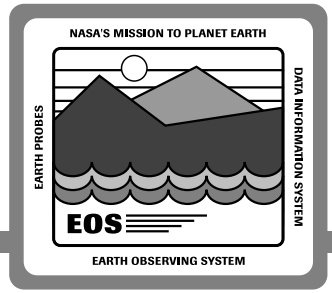
Support a fully interactive investigation capability



(OSU) HAIS needs to take a look at a much more interactive investigation model. This model uses frequent querying of the data from the SCF, interspersed with computational activity / requests. The interaction is quite fine grained, inconsistent with a coarse product ordering / generation capability.

(OSU) The SRR focused on data volumes. An investigation of data rates was missing. There was no emphasis on SCF-DAAC interaction. There was little analysis of how significant changes in network bandwidth might affect the way scientists work (i.e. the paradigm shift). The desktop must be considered a PART of EOSDIS that evolves with (in fact pushes evolution of) the system. ...

Science Drivers

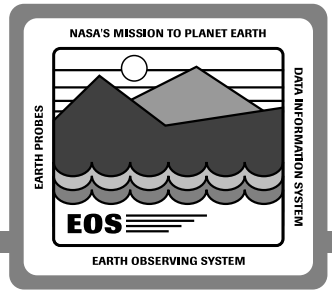


Support user-to-user collaboration

This driver reflects the scientists' desires for a streamlined environment in which data is readily available for interactive investigations. "Interactions" can include simple display of images, more complex data manipulations and visualization, and algorithmic processing (e.g. to attempt to extract new features).

(Skole - UNH) Dave's "model" is consistent with a more distributed "plug and play" service architecture, in which contributing researchers "publish" their data and access methods for the rest of the world. These products, however, are maintained (and even generated) locally, in response to dynamic requests from the community. We talked a bit about such a service-oriented architecture, including the need for local control of security and resource consumption. One of the concepts we discussed is the idea of SCFs acting as mini-PGS's, both ingesting and distributing data (to other SCFs).

Science Drivers

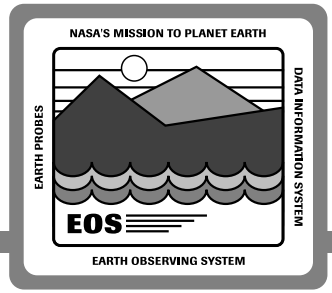


Facilitate an efficient search and "access" paradigm

The "search and order" paradigm is potentially too heavyweight. A lighter weight "search and access" paradigm should be employed, in which, once objects have been identified through specification in a search operation, they can simply be accessed (i.e. passed to an application, "opened", etc.).

(Consensus) Concern was voiced at SRR regarding the centralized, "heavyweight" nature of the product "ordering" scenario. Belief in the interdisciplinary science community is that data "access", as opposed to "ordering" is the logical end to the search process. The Andrew File System was suggested as a model for how to provide users with "access" to a global object space once the desired object has been identified.

Science Drivers

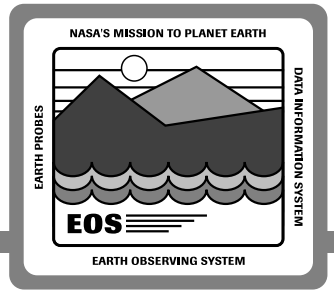


Support an information-rich data pyramid

One of the key characteristics the system should possess is the ability to retain data "lineage", or "heritage". Data heritage is the bulk of information that describes how a particular data object (it may be a product, derived metadata, etc.) was generated. It might include original source information (platform, sensor, date, time), plus any additional information used in processing (e.g. specific versions of calibration files), including algorithms, additional input datasets, etc. A rich data heritage in the "data pyramid" supports fundamental research questions about data and its sources, and allows scientists to explore the effects of alternative models in their work.

Science Drives

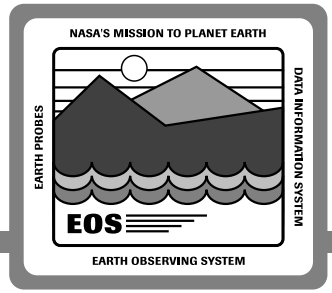
Support an information-rich data pyramid (Continued)



(Aber/Martin - UNH) ...it is important to be able to retrieve the raw data that had gone into other peoples published results. They would like to investigate the differences that different processing makes to the results. They said that different processing could (e.g., for atmospheric distortion) change estimate of the bottom line (e.g., forest canopy coverage) by as much as 50%. This potential for large differences points out the need to carefully document the transformations that are made to the data.

(OSU) ...one of the issues with current visualization tools the problem of "heritage tracing". The tools provide outstanding presentation and manipulation facilities, but lose track of the scientific basis ("validity") for the data being visualized.

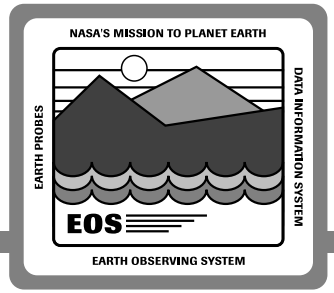
Science Drivers



Support the integration of independent investigator tools

- **ECS should provide the infrastructure and interoperability services to support researchers in using their own tools. This approach leverages the efforts that have gone into developing a rich set of existing investigative tools, and encourages future improvements in a competitive environment. This approach is key to encouraging scientific creativity in developing tools appropriate to the research task at hand. While ECS may wish to provide some simple set of basic analysis and visualization tools, these should be viewed as an optional set that can be replaced by more powerful or science specific tools.**

Support the integration of independent investigator tools (Continued)

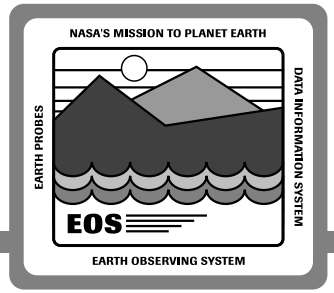


(Moore - UNH) Tools themes:

- Research is going on in heterogeneous environments (SCFs)
- The tools being used are NOT the same throughout the community
- The project should NOT be developing a single monolithic toolkit
- Rather, it should be building a common core around interprocess communications (DCE) and file transfer to support existing and future environments

(OSU) They use an extensive array of tools to support visualization and animation. These are clearly key components in the SCF of the near future (and the present!).

Science Drivers



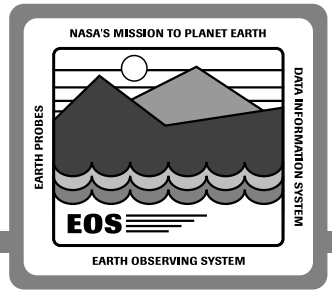
Provide distributed administration and control

- **The centralized SMC functionality and authority presented at SRR needs to be replaced by a decentralized, more autonomous entity. Granting of access privileges, especially to local facilities that may be part of ECS, should be done at the local level.**

(Emery) The system architecture presented at SRR was a centralized architecture that will not work. Concerns include:

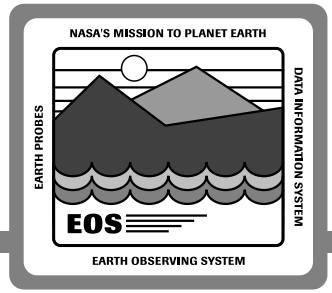
- the reliance on sponsoring "institutions". The system needs to deal with researchers and research groups as individuals.
- the SMC is a potential system bottleneck and single point of failure. Access to otherwise available data might be prevented by a problem within the SMC (e.g. maintenance). Distribute and replicate the SMC functionality to prevent this.

Summary of Science Drivers and Related ECS Analysis and Response



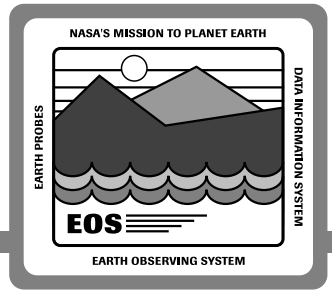
Science Driver	ECS Analysis and Response
• Support a dynamic product life cycle and easily extensible product set	• Extended service provider model (arch) • User/Data model • *GCDIS/UserDIS study*
• Support a fully interactive investigation capability	• Integrated data search and analysis as part of *Content-based search study*

Summary of Science Drivers and Related ECS Analysis and Response



Science Driver	ECS Analysis and Response
• Support user-to-user collaboration	• *GCDIS/UserDIS study* • Universal Reference (arch) • Interconnection Architecture Trade Study
• Facilitate an efficient search and "access" paradigm	• Universal Reference (arch) • *Content-based search study* • *Data distribution study* White paper and prototype • Service Routing Trade Study
• Support an information-rich data pyramid	• "Seamless view of data" (arch) • Data Server Architecture (arch)
• Support the integration of independent investigator tools	• Toolkit analysis and refinement (planned)
• Provide distributed administration and control	• Distributed nature of the new architecture • System Management Distribution Trade

Technology Drivers

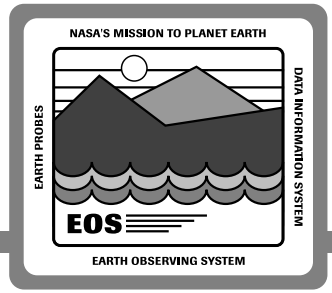


Technology drivers: **Technology advances that will enable ECS to provide increasing functionality to the users throughout the life of the system**

Advances in information technology software

- **Distributed database management**
- **Object oriented database management**
- **Extended relational database management**
- **Spatial and text data management**
- **Hierarchical storage management**

Technology Drivers
(Continued)

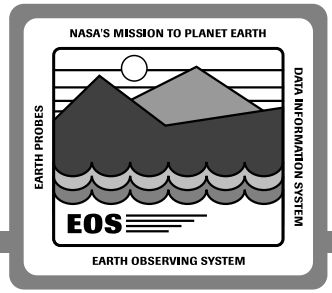


Advances in communications capability (Gigabit links)

Advances in processing capacity
(MPPs, proliferation, high performance workstations "farms")

Advances in multi-media technologies
(collaboration environments, videoconferencing)

Summary of Advanced Technology Drivers and Related ECS Analysis Response



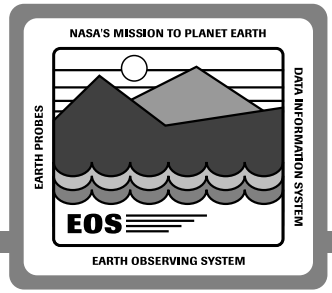
Technology Driver

- **Advances in information technology software**

ECS Analysis and Response

- **Sequoia 2000**
- **Prototyping of Object Oriented and Extended Relational DBMS (Study in progress, prototype planned for '94)**
- **Operating system standards**
- **Interconnection architecture trade study**
- **Evaluation of Standards (OMG, etc.)**
- **Prototyping of DCE**
- **Study and prototyping of Ellery Open Systems and Object Management Group**

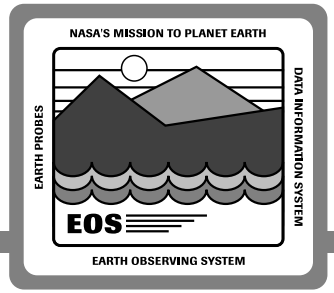
Summary of Advanced Technology Drivers and Related ECS Analysis Response (Continued)



- Advances in communications capability (Gigabit links)

- Extension of global naming and directory service
- NIIT/EDS (planned)
- Interconnection Architecture Trade Study
- Internet characterization study

Summary of Advanced Technology Drivers and Related ECS Analysis Response (Continued)



- **Advances in processing capacity**
  - **High end computing requirements (tall poles) study**
  - **Parallel and distributed computing testbed**
- **Advances in multi-media capability**
  - **GCDIS/UserDIS study**
  - **Interconnection Architecture Trade Study**