A Vision for EOL's new Computing, Data and Software (CDS) Section – Mike Daniels

Background and Context

Computing, software, data services and cyberinfrastructure have seen rapid and widespread changes over the past decade. In today's data-rich and highly networked environment, increased collaboration within the EOL, UCAR and the larger geosciences community is a necessity. Software engineering has played an increasing role in our measurement platforms, as there has been a trend toward more off-the-shelf components which have required extensive software development and integration. In a resource-limited environment, and to meet the challenges imposed by an emphasis on multi-disciplinary science, the EOL must design, build and maintain a cyberinfrastructure domain for its observing systems that will blend with the structure at UCAR and the geosciences community at large.

Mission

The Computing, Data and Software Section is responsible for developing and maintaining the integrity, capability, flexibility and currency of the computing infrastructure, data services, collaborative tools and software of the EOL, guided by and in service of our community.

Future Directions

- **Listen to our community through regular interactions related to data and software services.** The user survey from the EOL/JOSS software and data retreat gave us a hint as to what the community needs from us. We need to implement more regular interactions with them including the most feasible mechanisms to incorporate their feedback. Online surveys are one way, but we also need to have interactive discussions, workshops and participation in post-project analysis meetings on a regular basis. Input should be used in short- and long-term planning activities and in helping to determine our priorities.

- **Expand cyberinfrastructure for observing platforms and project data through partnerships and closer interactions with our community.** We cannot and should not do everything ourselves. Leveraging efforts within UCAR and our community is essential. UCAR wide data management is finally offering some payoffs in the way our data centers can be linked to other data centers and in the metadata search and discovery areas. Many higher-level tools such as IDL, MATLAB, NCL and the IDV are being used in the analysis of our data and we need to incorporate ways to produce datasets and data streams that are as compatible as possible. Working with new staff from JOSS/FODM, EOL data management must expand to encompass an entire project's dataset (not just EOL platforms). The next generation Zebra should be designed in partnership with other groups and include many more data sources (e.g. GIS, satellite and model output). Real-time forecast model assimilation will actually become a possibility through a combined effort with this community.

- **Improve the consistency of data quality processes in the Laboratory.** We need to structure the CDS in a way that facilitates the consistency aspects of data quality without separating the developers and data experts from their platforms. This will be a tricky personnel management effort, but the CDS needs to expand and be involved in more of these areas if we are to improve consistency and therefore our overall data quality. We need to look at how other organizations deal with QC and establish a process that can be used as a guideline and/or metric for all EOL platforms. Each platform needs a data quality expert who feels a sense of ownership for its data and who is dedicated to producing top quality results. We need to build laboratory buy-in for these improvements which will require extensive scientific guidance and a process to ensure that they continue to remain effective. Increased automation in the data QC process will become a necessity given the data volumes of the future. An EOL wide metadata database should be expanded to include more data sources, automating everything possible and generating metadata for a variety of QC/QA and documentation needs.

- **Maintain a flexible, secure and state-of-the-art computing infrastructure for the Laboratory, with frequent user guidance and input.** Computing and data capabilities as well as security issues will continue to increase. The staff in CDS needs to be empowered to implement technologies that will improve the infrastructure and keep pace with the changes in a proactive way, long before the impacts reach us. To increase communication, an EOL Computing User Group should be formed and led by non-CDS staff member. This group would be charged with improving communication and planning related to computing support issues and addressing the balance of user needs versus staffing resources. The group should meet quarterly and should contain representatives from the major staff groups.

- **Increase consistency among EOL software systems.** We need to reduce duplication of codes and algorithms and increase compatibility among our software systems. The RSF/SSSF merge and the HIAPER software development efforts have proven that the concept of a single software engineering group has widespread benefits. Through close interactions, our software engineering team can design more robust software whose components are shared among many EOL platforms. This has benefits from reducing the "one retirement away" syndrome to common user interfaces and better platform integration to the adoption of standard formats for data, data access and metadata. The work of our software engineers will be guided by scientific and engineering needs through the new EOL project management process. Our laboratory could serve as a repository for community developed software used in analyzing our data, linking researchers and reducing redundancies. A software engineer group can also be a good barometer for how well our Laboratory's matrix structure is working since its staff is likely to be involved in many projects. I propose that this group reside in the CDS since a major portion of its work is software support and I think more integration with data managers, computing systems and outside EOL efforts are needed which can ultimately improve automated data QC, real-time data access and cyberinfrastructure in our Laboratory.

- **Build a world-class web presence for our community**. With the exception of a few sections of well maintained content near the top levels, the current EOL web site may be described as chaotic at best. Using tools that connect to the metadata database, we should build an infrastructure that automates the creation and update of our data and instrumentation web pages. At the same time, this automation

will also bring more consistency to our web presence while at the same time reducing the staff time spent on this effort. An EOL-wide calendar of events would help the staff stay informed and platform or science specific online forums could build connections between scientists and students. Our experienced designers can build informational pages and web templates that present EOL topics in a clear and professional manner. These templates and training can ensure that EOL administrative staff is able to create web pages on their own, just as they do memos and other documents today. By using and relying on our web statistics we will insure that we are efficiently meeting our community's web needs by making commonly requested information easier to find and well organized.

- **Expand the use of EOL cyberinfrastructure in Education and Outreach opportunities.** There are many opportunities for the use of data and software in education and outreach and this should be expanded through partnerships with UCAR E&O and other groups. Case study datasets could be web accessible so that by clicking on a link, an integrated analysis tool is run containing all the available data sources so that may be explored interactively. Other E&O opportunities should be targeted at students of the atmospheric sciences in order to help them understand and correctly interpret data from our instrumentation. We should partner with scientists, QC experts and others to provide summary information and examples that can be used in this process. According to comments made by a UCAR Board of Trustees member, many of today's atmospheric science students lack an understanding of instrumentation issues and this is a key gap that must be addressed.

- **Build a virtual field operations center.** I can see some exciting links between the field catalog and the RDCC. We should combine these into one effort and build a linkage between the field catalog and the actual datasets themselves. There are many web and Internet technologies related to collaboration tools that should be explored, implemented and merged with existing tools. We can expand these tools to be a digital canvas that the scientists can use to post their interim results and findings, stay connected with participants who are working with the data and provide direction for the coordination of the project and its observing platforms. We should build on collaborations that have been initiated with UNIDATA, RAL, SCD and the UCAR Data Management Working Group (DMWG) in regard to data access, metadata and integrated displays. The next generation Zebra should be designed to be accessible across the Internet and we should build all EOL cyberinfrastructure with this remote access in mind. Real-time data from our observing platforms will be assimilated into models and this model output will be used to plan subsequent missions.

- **Advocate for and develop the careers of software engineers, systems administrators and data managers through effective EOL management representation**. These staff need mentoring, training and advancement opportunities beyond what has been available in the past. I feel there has been less emphasis on the importance of their activities and tools than what is appropriate since we rely so heavily on computing, software, data and networks today. I propose to make improvements by implementing a supportive organizational structure and insuring that these staff needs are addressed through advocacy at all levels of the EOL management process. As a metric, I would propose that we make SE and EE positions equitable in the Laboratory by FY09 and that we have

concrete plan for a 3 year upgrade cycle of all EOL desktop computing equipment in each facility. For the data managers, I would suggest that we lead the effort of UCAR groups to establish a formal Data Manager job profession at our institution. (Document continues below.)

Responsibilities, Resources and Proposed Organization

Computing Infrastructure

RESPONSIBILITIES

- In-house desktop & server support
- Network services
- Computing Security
- Field networks and Lab support
- Proactive planning of computing services and equipment upgrades

ADVISORY RESOURCES

- EOL Staff
- PIs

PARTNERSHIPS/COLLABORATION

- UCAR CSAC, ACCIS, NCAB, ITC
- Professional System Administration groups
- EOL Software Engineers
- UCAR System Administrators

STAFFING RESOURCES

- System Administrators (7)
- Students

Data Services

RESPONSIBILITIES

- EOL connections to UCAR & NSF cyberinfrastructure
- Internet data distribution & access
- End-to-end EOL metadata management
- Oversight of data quality coordination, tools and processes
- Project-wide data stewardship and services
- Centralized management of online and offline storage

ADVISORY RESOURCES

- EOL External Advisory Committee
- PIs
- EOL Scientists
- EOL Software Engineers
- UCAR Data Management Working Group

PARTNERSHIPS/COLLABORATION

- Geosciences data management community
- NSF Cyberinfrastructure
- UCAR data managers
- UCAR Labs & SE groups (ESSL, RAL, CISL, UNIDATA)

STAFFING RESOURCES

- Data Managers (5)
- Software Engineers (6)
- Students

Collaborative Technologies

RESPONSIBILITIES

- Web Engineering
- Network based collaboration
- Web data catalogs/browsers
- EOL web site maintenance and content management
- “Cyber” aspects of field coordination

ADVISORY RESOURCES

- EOL Staff
- PIs
- UCAR Web Advisory Group

PARTNERSHIPS/COLLABORATION

- Geosciences community
- PIs
- Professional web developers

STAFFING RESOURCES

- Web engineers (1)
- Software Engineers (2)
- Web Content/Designers (1)
- Students

Software Systems

RESPONSIBILITIES

- Development of data acquisition and display software for EOL’s observing platforms
- Internet based data streams
- Automated QC
- Real-time integrated displays and analysis tools
- Highly compatible datasets
- Software support

ADVISORY RESOURCES

- EOL Scientists and Engineers
- EOL External Advisory Committee
- PIs

PARTNERSHIPS/COLLABORATION

- UCAR SE assembly
- Professional Software Engineering groups
- EOL Engineers and Scientists
- UCAR Software Engineers
- PIs

STAFFING RESOURCES

- Software Engineers (7)
- Students

Staffing Resources

From earlier discussions, it is apparent that many of the EOL staff feel that the role of the CDS should be expanded to include areas such as field networks, data quality and software services. In many cases, to fulfill these responsibilities staff additions will be required. I would recommend that the EOL examine every new facility and section and determine the mix of staff resources, considering staffing demographics from the perspective of what will be required in the future. Depending on the scope of the expansion of its mission, it is estimated that the increases in staff could be in the ~18 FTE range over ATD/RDP levels. It is proposed that the software engineers be resident in the CDS so that the EOL can move to a more integrated approach to data acquisition, access, QC and display. Even with staff increases, it will be important for the CDS Manager and EOL to carefully manage staff resources and set priorities for CDS tasks. Administrative assistants should be appointed to the CDS and should assist in software licenses, contracts, budgets and other general administrative duties. We should factor in some percentage of staff salaries, perhaps up to 5% of the group's salary budget, to be used each year for either reclassifications or special accomplishment awards. Student assistant involvement in the CDS (and other EOL groups, for that matter) should be significantly expanded.

Advisory Resources

The CDS section has a responsibility to insure that its work is guided by internal and external sources. A regular schedule of, say, once/year for reviewing plans a progress and once every 3-5 years for reviewing strategic directions may be a good starting point. I would recommend that the advisory structure of the CDS section be formally identified and addressed and that the EOL Management lead this task.

Partnerships and Collaboration

The EOL must build partnerships with other groups within UCAR and the larger community as a whole. Under tight budgets and resources, we must work toward leveraging tools that have been developed by others in order to remain state-of-the-art. There are several efforts already in progress that we need to continue our involvement with, including the UCAR Data Portal, UNIDATA's IDV and real-time data transfer infrastructure, RAL's integrated software and satellite expertise and the UCAR GIS initiative just to name a few. New partnerships outside the atmospheric sciences community will also need to be developed as we move toward an environment of increased multi-disciplinary science.

Metrics

It takes a considerable amount of effort to design, implement, maintain and improve cyberinfrastructure, however, this work is often not recognized or understood by outside staff. In large part this is due to the fact that infrastructure is not visible (indeed, it should be transparent) unless it is not working. In order to both show the work being done and to insure that we are meeting the EOL's goals, I recommend that metrics be used to assess progress and achieve goals. A combination of subjective metrics (user surveys, opinions of outside expert panels) and objective metrics (uptime and downtime of servers, software bug reports, performance increases in computing, data and web infrastructure,

volume of data holdings and number of distributions, count of software download requests, help request statistics) should be used and reported on a regular basis.