A vision for end-to-end (E2E) data services:

"Unidata's vision calls for providing comprehensive, well-integrated and end-to-end data services for the geosciences. These include an array of functions for collecting, finding, and accessing data; data management tools for generating, cataloging, and exchanging metadata; and submitting or publishing, sharing, analyzing, visualizing, and integrating data."

What I have created is intended to be straw man- i.e., a first draft for beginning the discussion on what we mean by E2E data services and for providing comments, critique, and feedback so that we can collectively develop a common (read shared) vision for integrated, end-to-end data services. So by design, it is meant to be modified and further developed by all of you. For this concept paper, I have tried to focus on the What and not on the Why or How. Based on our strategic plan and the 5-year proposal, I hope everyone has some understanding of the Why. The How part will be addressed once we have developed a shared vision for E2E and are beginning to build "it". However, some of the How is mentioned briefly in this document as we won't be starting from scratch, but have a large collection of work already in place on which to build.

Overarching objective:

Development and deployment of a suite of data services that include many desired but customizable components and services that are well-integrated and enable us to realize the aforementioned vision is the overarching objective.

What end-to-end [or is it soup to nuts? :-)**] means to me?**

a) All stages of data life cycle beginning with observations:

Observations/Sensors → Ingest → Data collection systems → Data providers → Disseminate → Users (both end users and data archival systems)

b) From beginning till end of a workflow (LEAD example):

Observations → Ingest → Analysis/Assimilation → Prediction → Output → Dissemination → Users (both end users and data repositories)

This is an overly simple schematic of the workflow, but should be sufficient for this draft. One or more services will be needed at each stage, and they will need to be integrated both within a stage as well as across the stages.

Strategies, tactics and imperatives:

Integrated services do not imply building a monolithic system but a set of modular services that are configurable, flexible, extensible, and scalable. We need to think about a range of users and use cases (i.e., students, faculty, and researchers in

universities, data providers, scientists, use in field projects, use of services with real-time, case-study, and archived data, interactive and programmatic use, use by projects like LEAD and CADIS and portals like the ESG, CDP, etc.)

Integration can be achieved either via loosely or tightly coupled components and services.

Leverage as much as possible existing technologies, adopting and adapting them as much as possible.

We may not work on all of the desired functionalities or capabilities within the UPC, but we will need to think about what users need and find a way to facilitate many of the capabilities given our niche, competencies, and limited resources. If the architecture is modular and extensible, others will be able to add services and functionalities to it.

The $64K question: How best to build integrated, end-to-end data services upon and using the existing collection of software and with limited resources? Some but not all of the pieces are already well connected. *Incrementalism needs to be a key part of the tactic as we won't have the "luxury" of starting an effort ab initio and will need to leverage what is already developed.*

Data definition:

Includes:

Scientific data (binary, netCDF, ASCII, XML, ??)
Metadata (ASCII, XML, etc.)
DBMS
GIS Data (KML, shapefiles, etc.)
Derived products
Ancillary data objects (images, videos, documents (pdf, Word, html, ppt, etc.), and other information)

What we already have or are working on:

- TDS (including CDM, OPeNDAP, WCS, WMS, & cataloger)
- netCDF and related software
- RAMADDA
- ADDE
- UDUNITS
- LDM
- NOAAPORT ingest software
- Decoders for many types of data
- Next-generation LDM
- IDV, McIDAS, and GEMPAK, and possibly AWIPS II down the road

A list of possible (or needed) data services:

- Data collection service (routine ingest via LDM, FTP, etc.)
- Data submission service
- Metadata service for submitting, editing, and exchanging metadata
- Cataloging service
- Data discovery service
- Monitoring and notification service for new data, metadata, and products
- Data access services
- Data delivery/transport services, including copying/moving data to other servers and personal space and streaming data on demand
- Security and authentication services
- Subsetting service, including capability for progressive disclosure
- Aggregation services for data and metadata
- Services for CF conformance checking
- Decoding and data translation services
- Unit conversion services
- Visualization and product generation services
- Data fusion and data manipulation services (e.g., netCDF operator services)
- GIS services
- Output handling services

Note: I am using the term services loosely to describe components and functionalities. They don't necessarily have to be formal "web services".

Services can be invoked manually and programmatically by browsers, scripts, web services, workflows, different clients, etc.
(* We need to keep in mind that there are other popular clients beyond our own; e.g., GrADS, IDL, and Matlab.)

Data ingest/collection:

Enabled via a variety of mechanisms including:

- NOAAPort
- LDM
- Users, scripts, and workflows uploading
- Other sources and mechanisms

Data access and distribution:

- Both push and pull
- Interactive and programmatic (e.g., scripts)
- Invocation by web services (work flows)
- Access by different clients

- Subscription-based
- Authenticated/restricted
- Local and remote access
- Different protocols
- OGC services

Data Discovery:

- Search capability
- Browse capability
- Hooks for exchanging (or exporting) metadata, keywords, and tags

For example, show me what you have

- Temporally
- Spatially, and possibly using geographic names (gazetteer)
- By event/case (e.g., Katrina, Rita, etc.)
- By phenomenon (e.g., hurricane, tornado, etc.)
- Data type (observational, instrument, model, …)
- Show related items that have similar tags
- Others?

Integration capabilities:

- Different data types (e.g., feeds, obs. platforms, model output, including GIS information
- Different data formats
- Data on different projections
- Distributed data holdings
- Metadata addition
- Data manipulation and operation
- Scientific data with metadata content, documents, and other information

Not develop but provide hooks to:

- Collaboration tools
- Wikis
- Forums
- Blogs
- Chat and IM/SMS
- Social network applications (Facebook, Twitter, etc.)
- Notification by RSS, email, etc.
- Others

Beyond Unidata's scope (or strike zone), at least from my perspective:

- Data mining services
- Ontology services
- Brokering services
- Federation and mediation
- Provenance
- Curation and stewardship