AND Archives: Freeing Ourselves From the “Tyranny of the OR”

Ted Habermann
NOAA National Data Centers

This presentation is designed to be viewed as a PPT slide show.
Jim Collins (famous Boulder climber did first free assent of Genesis) and Jerry Porras did a study of Visionary Companies: premier institutions in their industries, widely admired by their peers and having a long track record of making a significant impact on the world around them. The key point is that a visionary company is an organization.

Identified characteristics of visionary companies through comparisons with comparable companies. One characteristic was:

Avoid the “Tyranny of the OR” by embracing the “Genius of the AND”.

## Tyranny of the OR

<table>
<thead>
<tr>
<th>purpose beyond profit</th>
<th>pragmatic pursuit of profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>a relatively fixed core ideology</td>
<td>vigorous change and movement</td>
</tr>
<tr>
<td>conservatism around the core</td>
<td>bold, committing, risky moves</td>
</tr>
<tr>
<td>clear vision and sense of direction</td>
<td>opportunistic groping and experimentation</td>
</tr>
<tr>
<td>Big Hairy Audacious Goals</td>
<td>incremental evolutionary progress</td>
</tr>
<tr>
<td>selection of managers steeped in the core</td>
<td>selection of managers that induce change</td>
</tr>
<tr>
<td>ideological control</td>
<td>operational autonomy</td>
</tr>
<tr>
<td>extremely tight culture (almost cult-like)</td>
<td>ability to change, move and adapt</td>
</tr>
<tr>
<td>investment for the long-term</td>
<td>demands for short-term performance</td>
</tr>
<tr>
<td>philosophical, visionary, futuristic</td>
<td>superb daily execution, “nuts and bolts”</td>
</tr>
<tr>
<td>organization aligned with a core ideology</td>
<td>organization adapted to its environment</td>
</tr>
</tbody>
</table>

**science information systems**

**geographic information systems**
THREDDS Data Server

HTTP Tomcat Server

Granule Metadata (Catalog.xml)

THREDDS Data Server (TDS)

NetCDF-Java library

• OPeNDAP
• HTTPServer
• OGC Web Coverage Service (WCS)

hostname.edu

Application

SIS AND GIS

CDM Datasets

Unidata’s Internet Data Distribution System
Data Processing Levels

- Level 0
  - Telemetry information, Swaths
  - Time and Scan Angle
  - Complex custom formats (bits)
  - Large volume
  - Radiance in instrument units
  - Complex and Hard

- Level 3 & 4
  - Grids
  - Latitude & Longitude
  - Standard formats (bytes)
  - Small volume
  - Sea Surface Temp °C
  - Simple and Easy

- POES Level 1b data
- 8km Level 2 SST
- NESDIS Products: 14, 50, 100km grids produced daily/weekly
- Most primitive useful form??
**NESDIS Level 2 Observations**

NESDIS (and Navy) Level 2 SST and Aerosol Observations are available via phone call / FTP arrangements with NCDC at present. These observations are in a custom format designed during the 1970's. The format has three major components: 5X5 spatial index, 1X1 spatial index, and the observations.

### Block Directory Record

<table>
<thead>
<tr>
<th>20 byte header</th>
<th>Block 1 Start Rec. #</th>
<th>Block 2 Start Rec. #</th>
<th>Block 3 Start Rec. #</th>
<th>...</th>
<th>Block 2592 Start Rec. #</th>
<th>Blanks</th>
</tr>
</thead>
</table>

### Observation Data Record

<table>
<thead>
<tr>
<th>Rec #</th>
<th>Block #</th>
<th>Extent #</th>
<th>Next Extent</th>
<th>Other Miscellaneous Stuff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subblock 1</td>
<td>Subblock 2</td>
<td>Subblock 3</td>
<td>...</td>
<td>Subblock 25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Start</th>
<th>End</th>
<th>Start</th>
<th>End</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
</table>

### Observation Unit

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Date / Time</th>
<th>Location</th>
<th>Observation</th>
<th>Other Miscellaneous Stuff</th>
</tr>
</thead>
</table>

**Observations**
Spatial Sorting and Indexing Point Data

Over the last decade commercial databases have developed the built-in capability to do this kind of spatial indexing. They bring many other capabilities to the table as well.

Satellite Data as points: Andy Pursch, Scott Shipley and someone @ NESDIS
OAIS Ingest Functions

Ingest Functions

Receive Submission
- SIP, AIP for Audit
- Data Formatting and Doc Std.
- SIP
- (Updated) SIP
- QA Results

Quality Assurance
- SIP
- Receipt confirmation
- Resubmit request

Generate AIP
- SIP
- Audit Report
- Descriptive Info

Co-ordinate Updates
- AIP
- Storage Request, AIP

Generate Descriptive Information
- AIP
- Report

PRODUCER

Archival Storage

DATA MANAGEMENT

Report Request
- Descriptive Info
- Database Update Request

Report
- Request
- Request
Archive Process Evolution

Heterogeneous Format Dependent Tools

Present Archive

Users

Designated Community

Homogeneous Data and Metadata, Standard Tools

Standard Metadata

Rich Granule Inventory

Standard Products

Future Archive
Step 1: Migrate the observations from a custom file format into a standard spatial database.

Step 2: Output a standard file format from the database.
A pipeline provides a description of a sequence of data processing tasks. The NGDC data processing pipeline provides a set of pipeline utilities designed around work queues that run in parallel to sequentially process data objects. The pipeline is an open source project hosted in the Jakarta Commons Sandbox (http://jakarta.apache.org/commons/sandbox/pipeline/).

Processing steps are specified as a series of stages in an XML configuration file.
SST Ingest Processing

Stage 1. Find Matching Files
Stage 2. Avoid Duplicate Processing
Stage 3. Read Data / Create Spatial Objects
Stage 4. Write Thinned Layer (10%) to DB & CDM
Stage 5. Write Complete Layer to DB & CDM
Stage 6. Create Summary (Grid) Table to DB & CDM
Stage 7. Create Rich Inventory Record
Integrated Visualization (GIS)
Partnership?

NOAA is a very different kind of organization than Unidata, but there are good signs:

NOAA Data Management Integration Team (DMIT) voted “Support for Common Data Model” as the #1 recommendation to IOOS for work that is consistent with the NOAA GEO-Integrated Data Environment Plan.

10 NOAA people attended Unidata training. 8 CLASS developers and others attending HDF Conference.
Formats and Products

Number of Products

Number of Formats

Sustainable?
Format Evolution

The graph illustrates the evolution of formats over time, categorized by producer-driven and user-driven approaches. The x-axis represents time, with producer-driven formats on the left and user-driven formats on the right. The y-axis shows the number of formats.

Key features include:
- **Producers**: A blue line indicating a steady decline over time.
- **Archive**: A red line showing a drop-off, peaking in the middle.
- **Users**: A black line indicating a significant decrease towards the right end, reflecting user-driven trends.

The graph visually captures how formats evolve in response to changes in the market dynamics, transitioning from producer to user-driven preferences.
Common Data Model

Scientific Datatypes
- Point
- Trajectory
- Station
- Radial
- Grid
- Swath

Coordinate Systems

Data Access

Open Geospatial Consortium Simple Features
Simple Features Spec

The Simple Feature Specification application programming interfaces (APIs) provide for publishing, storage, access, and simple operations on Simple Features (point, line, polygon, multi-point, etc). The purpose of these specifications is to describe interfaces to allow GIS software engineers to develop applications that expose functionality required to access and manipulate geospatial information comprising features with 'simple' geometry using different technologies.
The Rich Inventory Concept

Very similar to “file content metadata” at NCAR
1. Files come to CLASS and filename metadata is ingested into inventory.
2. Fileheader metadata is stored and is not available to data discovery system.
3. Descriptive Statistics are not calculated.
4. Users need to develop their own data discovery systems.
1. Files come to CLASS
2. Filename and fileheader metadata are added to inventory.
3. Descriptive Statistics are calculated and added to inventory.
4. All metadata is available to the data discovery system and users get the data they need without secondary data discovery.
Segment Model

- Constant (Static)
- Slow Variation (Quasi-static)
- Fast Variation (Dynamic)

Time (File Number)
Metadata Ingest

Create segment

New value?

Add to last segment

File

raw values

sum(x), sum(x^2), mean, std, count
Automated Observing System Ingest

Geospatial Database

Rich Inventory

Calculate simple statistics (SQL)

Pipelines

MADIS

ARGO

HADS

TABLE

TABLE

TABLE

HADS

ARGO

MADIS

Geospatial Database
HADS Network Monitoring
Algorithm Change: Aerosol
Hi Ted,
Dr. Ignatov and I did some digging and this is the result. Sasha's conclusion is the most pertinent info we could find from logs or email archives. Here it is:

Hi John,
i checked my 2002 email archives, and here is what i found out:

it appears that the current 3rd generation aerosol algorithm was implemented into operations around Oct-Nov 2002 time frame. cannot say more precisely, as all email correspondence i am looking at, talks about this indirectly. (maybe it's what Steve refers to as the Phase II aerosol-SST algorithm.) At the same time, Steve had implemented quite a few other changes fixing data bugs and formats: view angle problem in AEROBS, increased digitization in all channel's reflectances and AODs, etc.

The jump in AOD1 is deemed due to introducing 3rd generation algorithm, which replaced the 2nd generation. The new numbers (~0.08) look more realistic than the previous ones (~0.05 or so). The changes seen in the data is close to the expected effect of this change. the 3rd gen alg takes into account the exact spectral response of N16 AVHRR, whereas the 2nd gen was using a generic set of LUTs for all AVHRRs ("one size fits all").

hopefully this settles the issue..
cheers, sasha
1. Product generation algorithms write all metadata to inventory directly instead of file headers.
2. Files are archived somewhere with pointers from Inventory.
3. Users get the data they need from distributed system without secondary data discovery.