Provenance in Observational Solar Physics Data Pipelines

Peter Fox (HAO/ESSL/NCAR)
Deborah McGuinness (RPI)
Jose Garcia, Stephan Zednik (HAO/ESSL/NCAR)
Thomas Zurbuchen (Michigan)
Paulo Pinheiro da Silva (UTEP)
Funded by NSF/SDCI - OCI-0721943
Problem definition

- Data is coming in faster, in greater volumes and outstripping our ability to perform adequate quality control.

- Data is being used in new ways and we frequently do not have sufficient information on what happened to the data along the processing stages to determine if it is suitable for a use we did not envision.

- We often fail to capture, represent and propagate manually generated information that need to go with the data flows.

- Each time we develop a new instrument, we develop a new data ingest procedure and collect different metadata and organize it differently. It is then hard to use with previous projects.

- The task of event determination and feature classification is onerous and we don't do it until after we get the data.
Welcome to the Mauna Loa Solar Observatory (MLSO) Website. The MLSO, operated by the High Altitude Observatory in Boulder Colorado, houses several instruments designed to observe the sun at many different wavelengths. The MLSO instruments provide observations needed to understand the sun’s continuous release of plasma and energy into interplanetary space.

**ACOS**
Advanced Coronal Observing System. A suite of instruments designed to observe the solar atmosphere at a variety of heights. Includes Chromospheric Helium Imaging Photometer (CHIP, 1083 nm), H-alpha prominence and solar disk monitor (PICS, 656.2 nm), and the Mk4 K-coronameter, which observes the white light K-corona from 1.12–2.79 solar radii.

**ECHO**
Experiment for Coordinated Helioseismic Observations. A network of two instruments which observe solar oscillations as seen in the radial velocity of the solar surface.

**PSPT**
Precision Solar Photometric Telescope. Observes the solar disk in three bandpasses: 605–610 nm (red), 408–412 nm (blue), and 393 nm (CaII K).

### Latest MLSO Images

- **ACOS Mark IV**
  - K-Corona
  - 700-800 nm
  - 16-Jan-2007
  - Watch Movie

- **ACOS PICS Limb**
  - H-Alpha Limb
  - 656.3 nm
  - 16-Jan-2007
  - Watch Movie

- **ACOS PICS Disk**
  - H-Alpha Disk
  - 656.3 nm
  - 16-Jan-2007
  - Watch Movie

- **ACOS CHIP**
  - Helium I
  - 1083 nm
  - 16-Jan-2007
  - Movie (merged GIF)

- **PSPT CaII K**
  - PSPT Scale

- **PSPT Blue**
  - PSPT Scale

- **PSPT Red**
  - PSPT Scale

- **Velocity Image**
  - 1517:33
Use cases

- Determine which flat field calibration was applied to the image taken on January, 26, 2005 around 2100UT by the ACOS Mark IV polarimeter.
- Which flat-field algorithm was applied to the set of images taken during the period November 1, 2004 to February 28, 2005?
- How many different data product types can be generated from the ACOS CHIP instrument?
- What images comprised the flat field calibration image used on January 26, 2007 for all ACOS CHIP images?
- What processing steps were completed to obtain the ACOS PICS limb image of the day for January 26, 2005?
- Who (person or program) added the comments to the science data file for the best vignetted, rectangular polarization brightness image from January, 26, 2005 1849:09UT taken by the ACOS Mark IV polarimeter?
- What was the cloud cover and atmospheric seeing conditions during the local morning of January 26, 2005 at MLSO?
- Find all good images on March 21, 2008.
- Why are the quick look images from March 21, 2008, 1900UT missing?
- Why does this image look bad?
Provenance

• Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility.
• Determine which flat field calibration was applied to the image taken on January 26, 2005 around 2100UT by the ACOS Mark IV polarimeter.

• Which flat-field algorithm was applied to the set of images taken during the period November 1, 2004 to February 28, 2005.

• How many different data product types can be generated from the ACOS CHIP instrument?

• What images comprised the flat field calibration image used on January 26, 2007 for all ACOS Mark IV polarimeter images. 

• What processing steps were completed to obtain the ACOS PICS limb image of the day for January 26, 2005.

• Who (person or program) added the comments to the science data file for the best vignetted, rectangular polarization brightness image from January, 26, 2005 1849:09UT taken by the ACOS Mark IV polarimeter.

• What was the cloud cover and atmospheric seeing conditions during the local morning of January 26, 2005 at MLSO.

• Why does this image look bad?
Building blocks

• The proof markup language (PML) provides an interlingua for capturing the information agents need to understand results and to justify why they should believe the results.

• The Inference Web toolkit provides a suite of tools for manipulating, presenting, summarizing, analyzing, and searching PML in efforts to provide a set of tools that will let end users understand information and its derivation, thereby facilitating trust in and reuse of information.

• Capturing data quality, event, and feature detection within a suitable ontology -> VSTO
Yasukawa: Computer crash

Yasukawa: Rain, cloud
SPCDIS Search

Option: 

Query: quicklook from March 3 0:00 to March 3 23:59

Results 1 to 10 of 60 for 'quicklook from March 3 0:00 to March 3 23:59'

<table>
<thead>
<tr>
<th>label</th>
<th>type</th>
<th>source</th>
<th>Date &amp; Time</th>
<th>browse</th>
</tr>
</thead>
<tbody>
<tr>
<td>quicklook1.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 10:00</td>
<td>image</td>
</tr>
<tr>
<td>quicklook2.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 10:15</td>
<td>image</td>
</tr>
<tr>
<td>quicklook3.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 10:30</td>
<td>image</td>
</tr>
<tr>
<td>quicklook4.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 10:45</td>
<td>image</td>
</tr>
<tr>
<td>quicklook5.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 11:00</td>
<td>image</td>
</tr>
<tr>
<td>quicklook6.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 11:15</td>
<td>image</td>
</tr>
<tr>
<td>quicklook7.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 11:30</td>
<td>image</td>
</tr>
<tr>
<td>quicklook8.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 11:45</td>
<td>image</td>
</tr>
<tr>
<td>quicklook9.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 12:00</td>
<td>image</td>
</tr>
<tr>
<td>quicklook10.gif</td>
<td>...</td>
<td>...</td>
<td>March 3, 2008, 12:15</td>
<td>image</td>
</tr>
</tbody>
</table>
Summary

• Documenting and instrumenting MLSO data analysis pipeline
• Creating PML instances for integrating observer logs, quick look images, science FITS files
• Indexing these for smart search with constraints
• Answering several of the use cases
• Next steps: engineering data look up, science product use cases