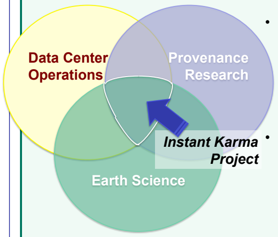
Instant Karma Status Update: Provenance at the AMSR-E SIPS

Helen Conover[1], Beth Plale[2], Mehmet Aktas[2], Bruce Beaumont[1], Dawn Conway[1], Sara Graves[1], Scott Jensen[2], Harsh Joshi[2],
Ajinkya Kulkarni[1], Yuan Luo[2], Robert Ping[2], Prajakta Purohit[2], Rahul Ramachandran[1], Kathryn Regner[1], Cara Stein[1]
[1]University of Alabama in Huntsville [2]Indiana University

Approach

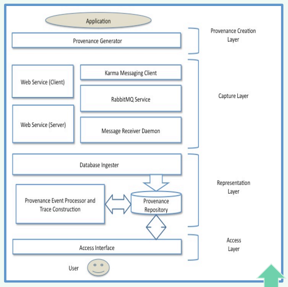


- Collaboration among
  - AMSR-E SIPS (MSFC Earth Science Office and UAHuntsville ITSC)
  - Provenance researchers at Indiana University's Data to Insight Center
  - AMSR-E Sea Ice science team (GSFC)
- Primary goal is to improve the collection, preservation, utility and dissemination of ***provenance information*** within the NASA Earth Science community
  - Using Karma provenance tool
  - Initial focus on Sea Ice processing

The Instant Karma project will integrate Karma, a provenance collection and representation tool developed at Indiana University, into the AMSR-E Science Investigator-led Processing System (SIPS) production environment, managed jointly by NASA/MSFC and UAHuntsville. The AMSR-E SIPS generates Level 2 and Level 3 data products from AMSR-E observations. An initial focus on Sea Ice processing will allow the project to engage the Sea Ice science team and user community in customizing Karma for NASA science data.

Karma Provenance Collection Tools

- Efficient and lightweight tools that support provenance collection, representation, and use
- Modular and programmable
  - Support diverse workflow architectures that consist of web services, java classes, message bus listeners



Karma logical architecture

- Capture provenance in streaming workflows
  - No need to know workflow structure in advance
- Support interoperability
  - Implement Open Provenance Model (OPM) v1.1* to represent provenance graph (access interoperability)
  - OPM enables provenance information exchange with other OPM-compliant tools
- Recent redesign of internal database schema and data structures represents Earth science relevant provenance more efficiently

* http://eprints.ecs.soton.ac.uk/16148/1/opm-v1.01.pdf

Provenance and Context Information

Lots of information already available, but scattered across multiple locations

- Processing system configuration
- Dataset and file level metadata
- Processing history information
- Quality assurance information
- Software documentation (e.g., algorithm theoretical basis documents, release notes)
- Data documentation (e.g., guide documents, README files)

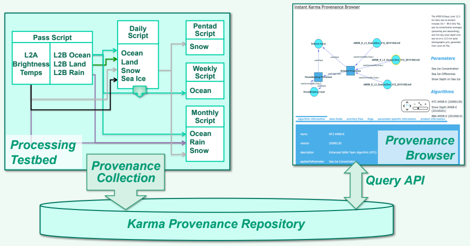
Instant Karma project aims to collate and organize information from multiple sources

Provenance Collection and Storage

AMSR-E SIPS processing workflow for Sea Ice instrumented in the testbed environment.

- Provenance information is captured in experiment run log files
- Log files are parsed to generate provenance notifications.
- These notifications are then imported into the Karma database.
- The Karma Service Query API is used to generate OPM-compatible XML graphs, each corresponding to a processing run.

Note that several of the services in the sea ice workflow are housekeeping and processing automation scripts, which are part of the processing workflows for other AMSR-E daily products.



Defining and Collecting Science-Relevant Provenance and Context

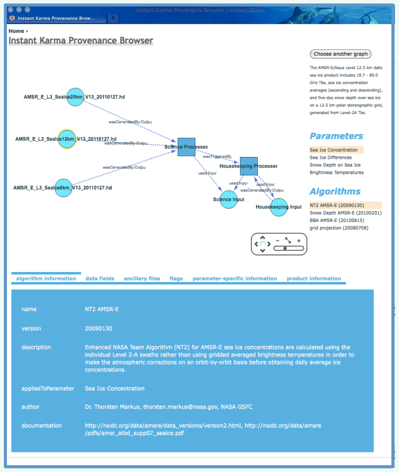
- Harvesting granule information from ECS metadata
- Also recording processing location associated with each data granule
- Working with AMSR-E Science Computing Facility to identify algorithm and data product information
  - Algorithm versions and descriptions
  - Parameters and data fields
  - Ancillary files
  - Flag values and explanations
  - Pointers to full documentation
- Defining how to harvest, transmit and display this information

Science Use Cases

- ✓ Browse provenance graphs : convey rich information about final data granule details [Use case 1]
  - Spatial location, time of observation, algorithms employed, input data and ancillary files
  - Provenance bundle to include pointers to relevant documentation
- ✓ Answer "Something isn't right" question [Use case 1 variant]
  - E.g., did not receive data for several days so snow melt mask may be inaccurate.
- Compare two data granules [Use case 2]
  - Query system to get list of provenance differences (e.g., versions of software, number and versions of input files)
- ✓ General provenance graph for a given science process, e.g., Sea Ice processing [Use case 3]
  - Current algorithms and versions, nominal number and versions of input files, pointers to relevant documentation
- Embed provenance information as annotations in HDF files

Browsing Provenance Information

- Interactive web application allows users to view the provenance graph for a specified data product
- Click on a node to display the full description of the product or process
- Trace full lineage of a data product by viewing the provenance information for each input file
- Access relevant information for the data product
  - Algorithm Theoretical Basis Document
  - README files
  - Product and inventory level metadata
- Uses Karma Service Query API to extract provenance graphs from the Karma provenance repository.


Instant Karma Provenance Browser | Instant Karma
Instant Karma Provenance Brow...
Home ›
Instant Karma Provenance Browser
Choose another graph
AMSR_E_L3_SeaIce25km_V13_20110127.hd
AMSR_E_L3_SeaIce12km_V13_20110127.hd
AMSR_E_L3_SeaIce6km_V13_20110127.hd
wasGeneratedBy:Outpu
Science Processes
wasTriggeredBy
Housekeeping Processes
usedInput
Science Input
Housekeeping Input
algorithm information
data fields
ancillary files
flags
parameter-specific information
product information

The AMSR-E/Aqua Level 12.5 km daily sea ice product includes 18.7 - 89.0 GHz Tbs, sea ice concentration averages (ascending and descending), and five-day snow depth over sea ice on a 12.5 km polar stereographic grid, generated from Level-2A Tbs.

Parameters

Sea Ice Concentration
Sea Ice Differences
Snow Depth on Sea Ice
Brightness Temperatures

Algorithms

NT2 AMSR-E (20090130)
Snow Depth AMSR-E (20100201)
BBA AMSR-E (20100615)
grid projection (20080708)

name	NT2 AMSR-E
version	20090130
description	Enhanced NASA Team Algorithm (NT2) for AMSR-E sea ice concentrations are calculated using the individual Level 2-A swaths rather than using gridded averaged brightness temperatures in order to make the atmospheric corrections on an orbit-by-orbit basis before obtaining daily average ice concentrations.
appliesToParameter	Sea Ice Concentration
author	Dr. Thorsten Markus, thorsten.markus@nasa.gov, NASA GSFC
documentation	http://nsidc.org/data/amsre/data_versions/version2.html, http://nsidc.org/data/amsre/pdfs/amsr_atbd_supp07_seaice.pdf

Browser prototype showing provenance graph and related information for generation of a daily 12.5 km Sea Ice product from AMSR-E Brightness Temperatures.

Acknowledgements: The Instant Karma project, funded under NASA ACCESS program, is a collaboration among NASA, UAHuntsville and Indiana University. The project team includes PI Michael Goodman (NASA/MSFC), Science Co-I Thorsten Markus (NASA/GSFC), Co-I's Helen Conover (UAHuntsville) and Beth Plale (IU) and their teams.

UAHuntsville
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
PERVASIVE TECHNOLOGY INSTITUTE