Scientific Workflows and the Sensor Web

GEOSS Future Products Workshop
Washington, Maryland, USA
26 March 2013

Derek Hohls
dhohls@csir.co.za

CSIR
our future through science

Outline

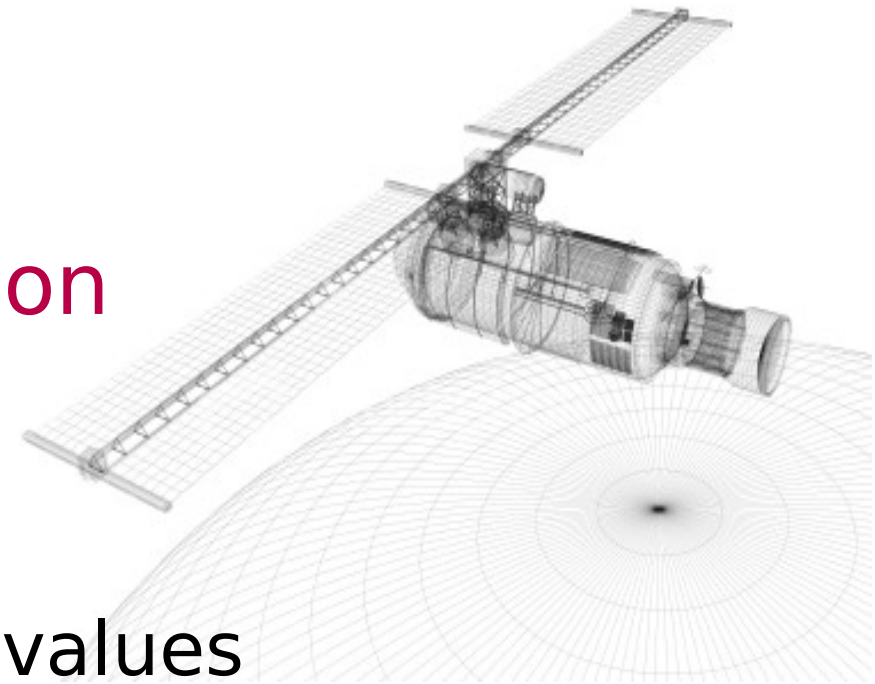
- Current challenges
- A proposed solution
- “In detail”
- Summary of progress
- Future directions

Conclusion

access to sensor web data
via scientific workflows
where aspects can be automated
can improve the process of “doing science”

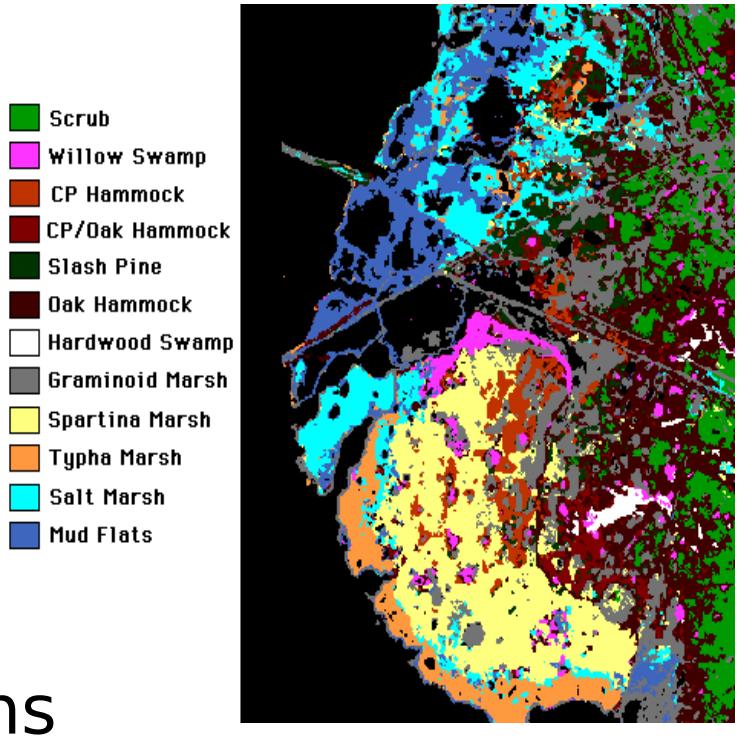
Current challenges: sensors

- Sensor discovery
- Sensor access
- Data transformation
  - Discretisation
  - Continuisation
  - Imputing missing values



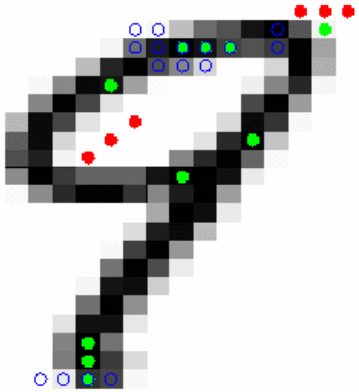
Current challenges: spatial

- Feature creation
  - FFT, feature extraction
- Data scrubbing
- Data integration
  - Incompatible data types
  - Co-ordinate transformations
  - Integration of temporal and spatial data into statistical and machine learning frameworks



Current challenges: processing

- Data exploration
  - Feature selection
  - Dimensionality reduction
  - Visualisation

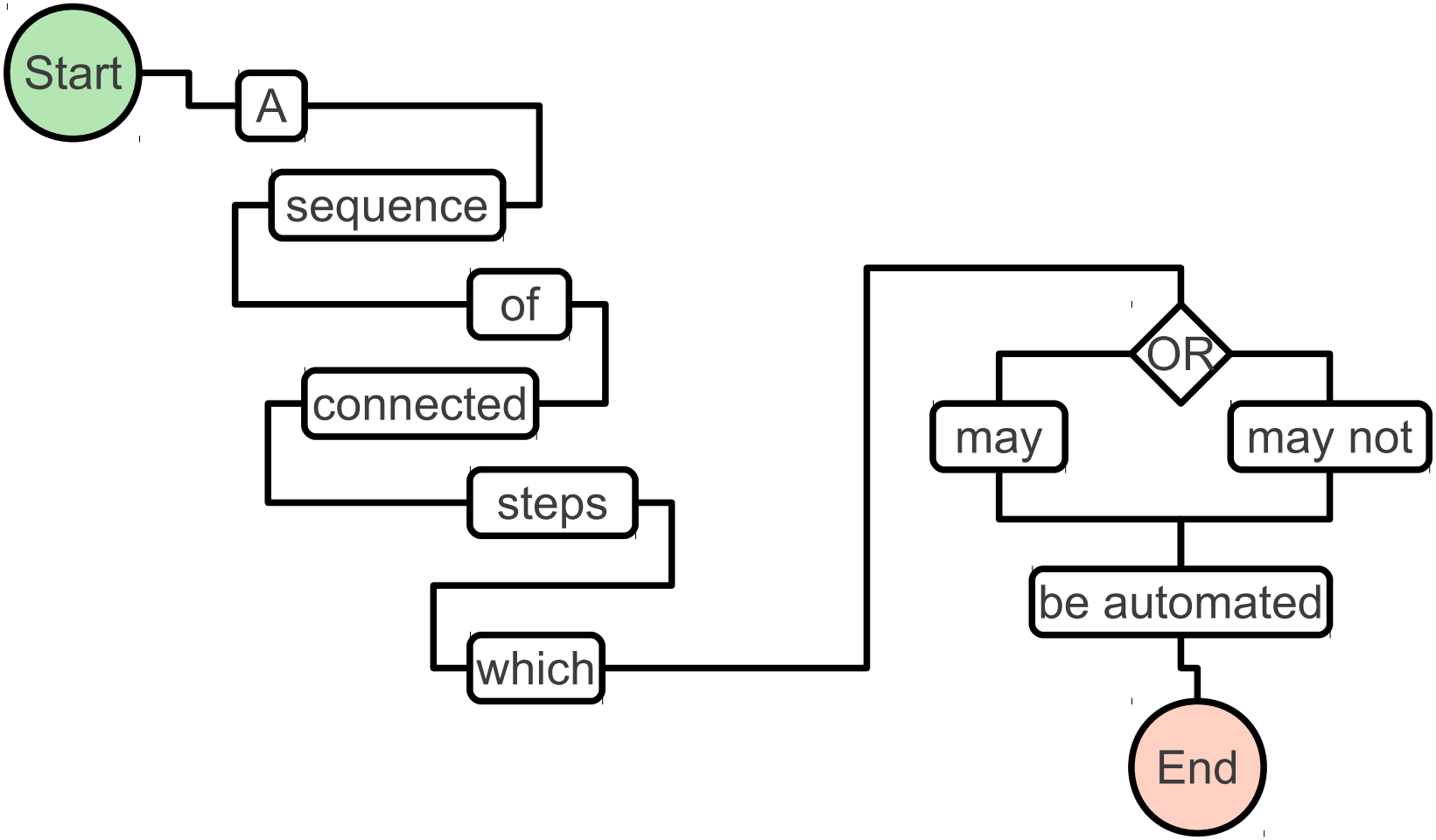


- Data process selection and chaining
- Distributing workload over HPC clusters or cloud computing resources
- Capture provenance to enable repeatibility

Solutions #1: Sensor Web … and the gap

- Sensor discovery ☺ CSW
- Sensor access ☺ SOS
- Data transformation ☺ SWE
- Feature creation ?
- Data scrubbing ?
- Data integration ?
- Data exploration ☺ CSW
- Data process selection ☺ WPS
- Distributed Computing ☺ SWE
- Provenance ?
- Repeatability ?

Solutions #2: Using a workflow?



Solutions #3: A scientific workflow

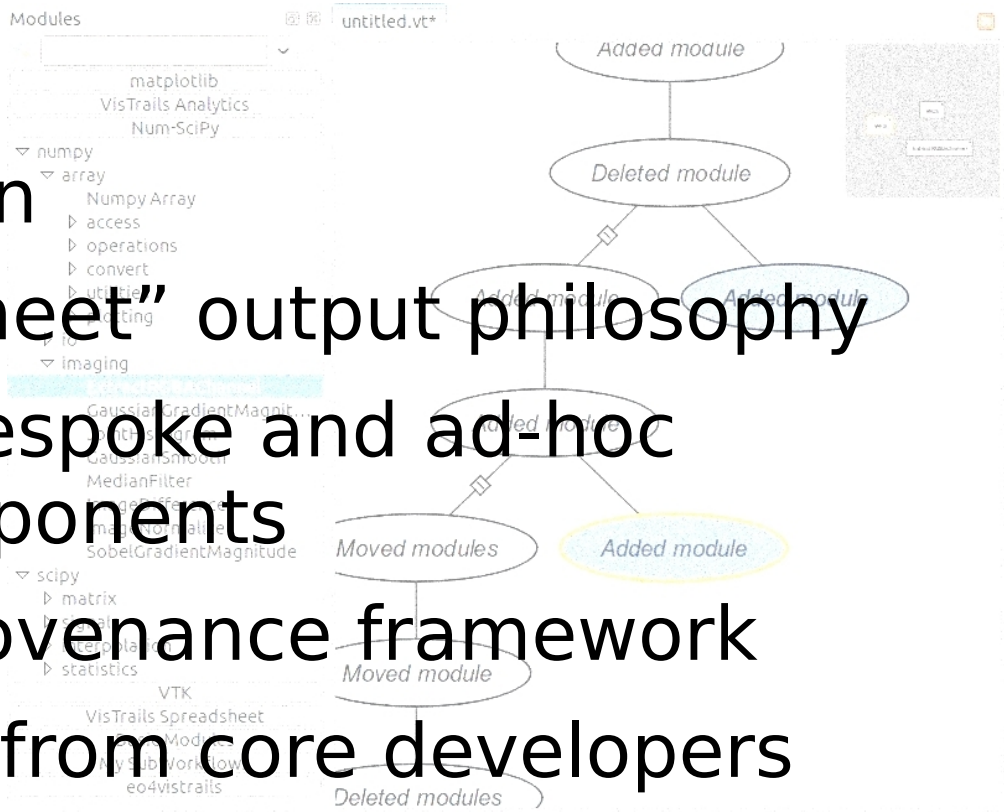
- Facilitate exploration
- Catalogue experiments (provenance)
- Enable repeatability
- Allow portability (process sharing)
- Provide domain specific tool-sets
  - Bio-Informatics
  - Physics

Solutions #3: Sensor Web + Scientific Workflows

- The gap:
  - No open source solution exists for a sensor web enabled “workbench”
- The offering:
  - EO4VisTrails provides spatial and temporal data access, data pre-processing and data analysis capabilities

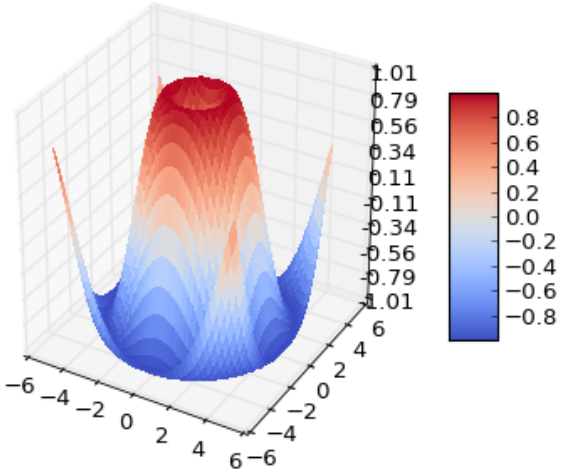
EO4VisTrails: Rationale #1

- Why VisTrails?
  - Built on Python
  - Has “spreadsheet” output philosophy
  - Allows both bespoke and ad-hoc workflow components
  - Embedded provenance framework
  - Good support from core developers



EO4VisTrails: Rationale #2

- Why Python?
  - A language designed for ease-of-use
  - Extensive scientific libraries
    - numpy, scipy, matplotlib
  - Wrappers for scripting
    - R, PySAL

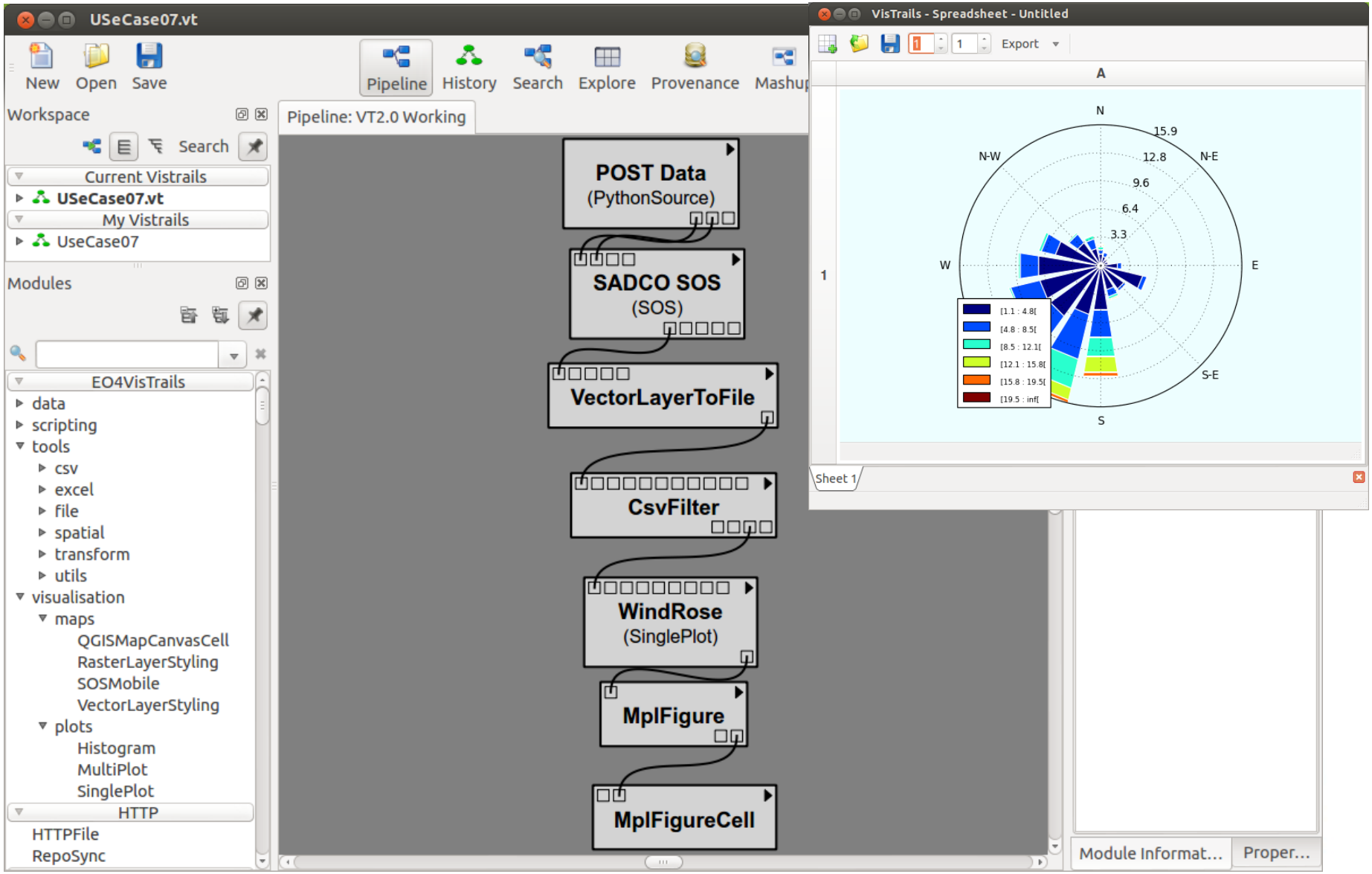


“In detail” EO4VisTrails Components

- Core OGC services
  - SOS
    - Data Access
    - Register Sensor
    - Insert Observation
  - WCS
  - WFS
  - WMS

Other Module Groups
Scripting wrappers
PostGIS access
Map display (QGIS API)
Excel manipulation
Remoting (RPyC)
OpenDAP/NetCDF
Pre-processing
Post-processing

Workflow #1: SOS → WindRose



Workflow #1: SOS Settings

VisTrails - Tools

Module Configuration

Service Metadata | SOS Specific Metadata | Bounding Coordinates | Temporal Bounds and Intervals

OGC Sensor Observation Service:

URL & Version: http://ict4eo.meraka.csir.co.za/sadcosos/sos.py 1.0.0 Fetch

Service Metadata

Service Identification

Service	OGC:SOS
Version	1.0.0
Title	SADCO Sensor Observation Service
Abstract	weather and oceanography monitoring network
Keywords	['AfriSpatial', 'CSIR', 'SADCO']
Fees	NONE
Access Constraints	NONE

Publisher Details

Provider Name	AfriSpatial
Provider URL	http://afrispatial.co.za
Contact Name	Gavin Fleming
Contact Position	Tech manager
Contact Role	None
Contact Organization	None
Contact Address	Box 436
Contact City	Franschhoek
Contact Region	Western Cape
Contact Postal Code	7690
Contact Country	South Africa
Contact Phone	0218620670
Contact Fax	0866164820
Contact Site	http://afrispatial.co.za
Contact Email	gavin@afrispatial.co.za
Contact Hours	None
Contact Instructions	None

CSIR

our future through science

Workflow #1: SOS MetaData

VisTrails - Tools

Module Configuration

Service Metadata | SOS Specific Metadata | Bounding Coordinates | Temporal Bounds and Intervals

Offerings

- watphy
- wet_data

Offering Details

Description	SADCO terrestrial weather station data
Bounding Box	Top Left X: - Top Left Y: -
	Bottom Right X: - Bottom Right Y: -
	SRS: -
Time	1963-03-31T23:00:00+02:00 to: 2010-05-31T21:50:00+02:00
Procedure	AWS_at_AB01
Response Format	text/xml;subtype='sensorML/1.0.0'
Response Mode	inline
Result Model	om:Observation
Observed Property	urn:ogc:def:property:x-sadco:0.1:wind_dir urn:ogc:def:property:x-sadco:0.1:wind_speed_ave urn:ogc:def:property:x-sadco:0.1:wind_speed_max
Feature of Interest	urn:ogc:object:feature:x-sadco:0.1:station:DB03
Time Limit?	
Spatial Delimiter?	
Request Type	GetObservation

Cancel OK

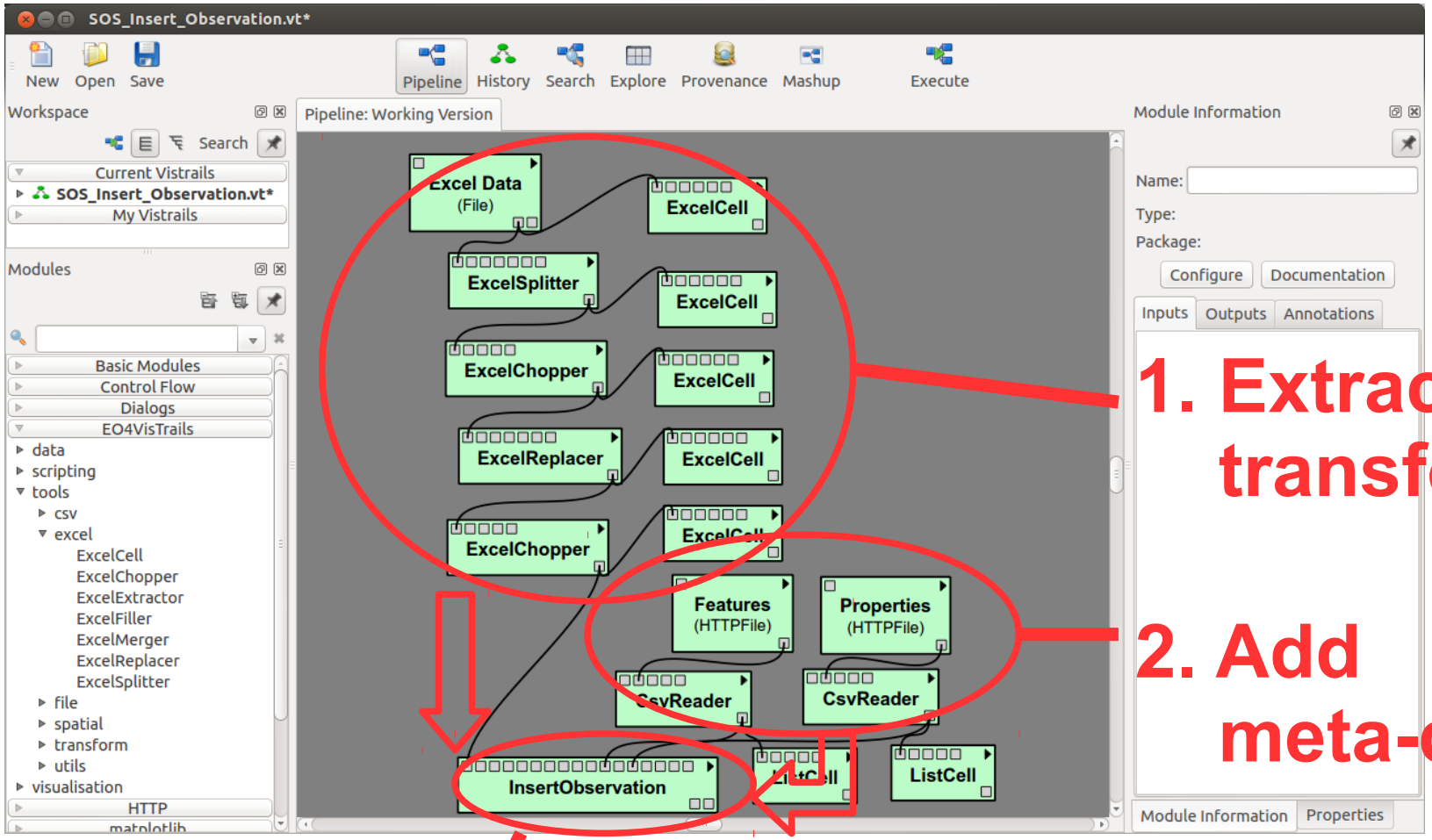
CSIR

our future through science

Workflow #2: Data for a SOS

- Excel stores data in an “unstructured” form
- Scientists have creative ways of arranging and labeling data
- Excel data cannot easily be processed & shared

Workflow #2: SOS Insert Observation



1. Extract & transform

2. Add meta-data

3. Insert data

Progress-to-date

- Workflows have demonstrated their usefulness in our sphere of work
- Individual modules have been demonstrated/used in projects
- Some modules are fully “operational”; others are “under development”

Future Directions #1

- Complete documentation & **packaging**
- Conduct **research** into other modules
  - PySAL for spatial analysis
  - Data access modules
- Grow the **community**
  - Server-based workflows – CrowdLabs
  - Client-focused module development e.g.
    - Carbon Flux Modelling

Fin

Thank you for your time!

Appendix I: Key Links

- Open Geospatial Consortium
  - http://www.opengeospatial.org
- VisTrails
  - http://www.vistrails.org
- EO4VisTrails
  - http://code.google.com/p/eo4vistrails/

Appendix II: Technologies

OGC®
Making location count.

scikits learn

RPy

GDAL

PyTables

NumPy

Appendix III: EO4VisTrails Architecture