Mind Your Metadata

Exploiting Semantics for Configuration, Adaptation, and Provenance in Scientific Workflows

Yolanda Gil
Pedro Szekely
Craig Knoblock
Varun Ratnakar
Shubham Gupta
Maria Muslea
Fabio Silva

Tom Harmon
Sandra Villamizar

UC Merced

USC/ISI
River Continuum vs Human Activities

- River continuum: natural inputs, reactive transport
- Human intervention: Agricultural, industrial, municipal
- What management practices help/hurt?
- Can we restore natural behavior?
Case Study

San Joaquin River to San Francisco Bay
UC Merced station
MST
Synoptic reach
San Joaquin River from Sierra Nevada
Merced River
CRS
NEW
Stream Metabolism Response to Human Disturbances

Pulse releases in the spring and fall to help the salmon runs
Stream Metabolism Response to Human Disturbances

Pulse releases in the spring and fall to help the salmon runs

... but how does this affect the ecology of the river?

... how about the effect of farmers?
Aquatic Photosynthesis

Models of gross primary production (GPP), community respiration (CR24)

Sensors
Aquatic Photosynthesis

Models of gross primary production (GPP), community respiration (CR24)

Sensors

Analyses must be fast to produce actionable information
Workflow

Tom Harmon
environmental systems
Vision: Automated & Fast
Reality: Difficult & Time Consuming
Current Method

Manual Data Preparation

Custom Scripts
Current Method

Multiple, separate tools
High learning costs
Ad hoc, by-hand movement of data & tool invocation
Data does not “flow” across tools
Our Approach

KARMA

Semantic Metadata

WINGS
Data Sources

- Water Quality Sensors (local)
- CDEC sensors (government)
- NOAA (government)

Processing Steps:
- Normalize
- Normalize
- Resample Hourly
- Clean
KARMA

[Tuchinda et al TWEB’11; Tuchinda et al IUI’08, IUI’07]
KARMA

[Tuchinda et al TWEB’11; Tuchinda et al IUI’08, IUI’07]
Data Import

![Data Import Interface](image)
Need to Clean Data

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>RIVER STAGE (feet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20100309</td>
<td>2300</td>
<td>52.68</td>
</tr>
<tr>
<td>20100309</td>
<td>2315</td>
<td>52.68</td>
</tr>
<tr>
<td>20100309</td>
<td>2320</td>
<td>52.68</td>
</tr>
</tbody>
</table>

CDEC

<table>
<thead>
<tr>
<th>timestamp</th>
<th>WXT510P</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-03-10 00:00:00</td>
<td>760</td>
</tr>
<tr>
<td>2010-03-10 00:15:00</td>
<td>760</td>
</tr>
</tbody>
</table>

HYDROLAB

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Temp</th>
<th>Cond</th>
</tr>
</thead>
<tbody>
<tr>
<td>03/09/2010</td>
<td>23:00</td>
<td>13.4</td>
<td>1181.00</td>
</tr>
<tr>
<td>03/09/2010</td>
<td>23:15</td>
<td>13.4</td>
<td>1179.00</td>
</tr>
</tbody>
</table>

Required Format
Need to Clean Data

60 Files for 1 month!

CDEC

HYDROLAB

Required Format
Data Cleaning with KARMA
Data Cleaning with KARMA

User provides example

KARMA generates cleaning rule
Need to Integrate All the Sources

Integrated Data Set
Joined by Date & Hour
### Integrated Dataset

<table>
<thead>
<tr>
<th>Station ID</th>
<th>forSite</th>
<th>startDate</th>
<th>forDate</th>
<th>String</th>
<th>String</th>
<th>Temp</th>
<th>Cond</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMN</td>
<td>03/10/2010</td>
<td>03/09/2010</td>
<td>23:00</td>
<td>13.4</td>
<td>1181.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMN</td>
<td>03/10/2010</td>
<td>03/09/2010</td>
<td>23:15</td>
<td>13.4</td>
<td>1179.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMN</td>
<td>03/10/2010</td>
<td>03/09/2010</td>
<td>23:30</td>
<td>13.4</td>
<td>1184.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMN</td>
<td>03/10/2010</td>
<td>03/09/2010</td>
<td>23:45</td>
<td>13.3</td>
<td>1185.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMN</td>
<td>03/10/2010</td>
<td>03/10/2010</td>
<td>00:00</td>
<td>13.3</td>
<td>1185.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMN</td>
<td>03/10/2010</td>
<td>03/10/2010</td>
<td>00:15</td>
<td>13.3</td>
<td>1182.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMN</td>
<td>03/10/2010</td>
<td>03/10/2010</td>
<td>00:30</td>
<td>13.3</td>
<td>1182.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
KARMA Generates Data Processing Script

importWSSource("CDEC Event Data","SMN","146","$1","$2");
setColumnName("CDEC Event Data","4","Date");
applyCleanRule("CDEC Event Data","Date","20100309","03/09/2010");
deleteColumnCommand("Sensor");
deleteColumnCommand("End Date");
setColumnName("CDEC Event Data","3","Time");
setColumnName("CDEC Event Data","4","Temp");
applyCleanRule("CDEC Event Data","Time","2300","23:00");
switchToEmptySourceTab(1);
importWSSource("CDEC Event Data","SMN","100","$1","$2");
setColumnName("CDEC Event Data1","4","Date");
applyCleanRule("CDEC Event Data1","Date","20100309","03/09/2010");
deleteColumnCommand("Sensor");
deleteColumnCommand("Start Date");
deleteColumnCommand("End Date");
setColumnName("CDEC Event Data1","2","Time");
setColumnName("CDEC Event Data1","3","Cond");
applyCleanRule("CDEC Event Data1","Time","2300","23:00");
switchToEmptySourceTab(2);
importWSSource("CDEC Event Data","SMN","1","$1","$2");
setColumnName("CDEC Event Data2","4","Date");
applyCleanRule("CDEC Event Data2","Date","20100309","03/09/2010");
deleteColumnCommand("Sensor");
deleteColumnCommand("Start Date");
deleteColumnCommand("End Date");
setColumnName("CDEC Event Data2","2","Time");
setColumnName("CDEC Event Data2","3","Depth");
applyCleanRule("CDEC Event Data2","Time","2300","23:00");
switchToEmptySourceTab(3);
importWSSource("CDEC Event Data","SMN","61","$1","$2");
setColumnName("CDEC Event Data3","4","Date");
applyCleanRule("CDEC Event Data3","Date","20100309","03/09/2010");
deleteColumnCommand("Sensor");
deleteColumnCommand("Start Date");
deleteColumnCommand("End Date");
setColumnName("CDEC Event Data3","2","Time");
setColumnName("CDEC Event Data3","3","DO");
applyCleanRule("CDEC Event Data3","Time","2300","23:00");
switchToEmptySourceTab(4);
importWSSource("CDEC Event Data","SMN","20","$1","$2");
setColumnName("CDEC Event Data4","4","Date");
applyCleanRule("CDEC Event Data4","Date","20100309","03/09/2010");
deleteColumnCommand("Sensor");
deleteColumnCommand("Start Date");
deleteColumnCommand("End Date");
setColumnName("CDEC Event Data4","2","Time");
setColumnName("CDEC Event Data4","3","Flow");
applyCleanRule("CDEC Event Data4","Time","2300","23:00");
switchToSourceTab(0);
join("CDEC Event Data0","CDEC Event Data1","Cond");
join("CDEC Event Data0","CDEC Event Data2","Depth");
join("CDEC Event Data0","CDEC Event Data3","DO");
join("CDEC Event Data0","CDEC Event Data4","Flow");
switchToSourceTab(5);
importCSVSource("data\CDEC_stations.csv");
importColumnFromCSV("Station ID","0","true");
importColumnFromCSV("Metadata","1","true");
importColumnFromCSV("Name","2","true");
importColumnFromCSV("Elevation","3","true");
importColumnFromCSV("Latitude","4","true");
importColumnFromCSV("Longitude","5","true");
switchToSourceTab(0);
join("CDEC Event Data0","CDEC_stations.csv","Latitude");
join("CDEC Event Data0","CDEC_stations.csv","Longitude");
publishToWS("WINGS Portal","TEST_CDEC_WEATHER_$3","CDEC Event Data0");
KARMA Generates Data Processing Script

ImportWSSource("CDEC - Event Data","SMN","146","$1","$2");
SetColumnName("CDEC - Event Data0","4","Date");
ApplyCleanRule("CDEC - Event Data0","Date","20100309","03/09/2010");
DeleteColumnCommand("Sensor");
DeleteColumnCommand("End Date");
SetColumnName("CDEC - Event Data0","3","Time");
SetColumnName("CDEC - Event Data0","4","Temp");
ApplyCleanRule("CDEC - Event Data0","Time","2300","23:00");
SwitchToEmptySourceTab(1);
ImportWSSource("CDEC - Event Data","SMN","100","$1","$2");
SetColumnName("CDEC - Event Data1","4","Date");
ApplyCleanRule("CDEC - Event Data1","Date","20100309","03/09/2010");
DeleteColumnCommand("Start Date");
DeleteColumnCommand("End Date");
SetColumnName("CDEC - Event Data1","2","Time");
SetColumnName("CDEC - Event Data1","3","Cond");
ApplyCleanRule("CDEC - Event Data1","Time","2300","23:00");
SwitchToEmptySourceTab(2);
ImportWSSource("CDEC - Event Data","SMN","1","$1","$2");
SetColumnName("CDEC - Event Data2","4","Date");
ApplyCleanRule("CDEC - Event Data2","Date","20100309","03/09/2010");
DeleteColumnCommand("Sensor");
DeleteColumnCommand("Start Date");
DeleteColumnCommand("End Date");
SetColumnName("CDEC - Event Data2","2","Time");
SetColumnName("CDEC - Event Data2","3","Depth");
ApplyCleanRule("CDEC - Event Data2","Time","2300","23:00");
SwitchToEmptySourceTab(3);
ImportWSSource("CDEC - Event Data","SMN","61","$1","$2");
SetColumnName("CDEC - Event Data3","4","Date");
ApplyCleanRule("CDEC - Event Data3","Date","20100309","03/09/2010");
DeleteColumnCommand("Sensor");
DeleteColumnCommand("Start Date");
DeleteColumnCommand("End Date");
SetColumnName("CDEC - Event Data3","2","Time");
SetColumnName("CDEC - Event Data3","3","DO");
ApplyCleanRule("CDEC - Event Data3","Time","2300","23:00");
SwitchToEmptySourceTab(4);
ImportWSSource("CDEC - Event Data","SMN","20","$1","$2");
SetColumnName("CDEC - Event Data4","4","Date");
ApplyCleanRule("CDEC - Event Data4","Date","20100309","03/09/2010");
DeleteColumnCommand("Sensor");
DeleteColumnCommand("Start Date");
DeleteColumnCommand("End Date");
SetColumnName("CDEC - Event Data4","2","Time");
SetColumnName("CDEC - Event Data4","3","Flow");
ApplyCleanRule("CDEC - Event Data4","Time","2300","23:00");
SwitchToSourceTab(0);
join("CDEC - Event Data0","CDEC - Event Data1","Cond");
join("CDEC - Event Data0","CDEC - Event Data2","Depth");
join("CDEC - Event Data0","CDEC - Event Data3","DO");
join("CDEC - Event Data0","CDEC - Event Data4","Flow");
SwitchToSourceTab(0);
ImportCSVSource("CDEC_stations.csv");
ImportColumnFromCSV("Station ID","0","true");
ImportColumnFromCSV("Metadata","1","true");
ImportColumnFromCSV("Name","2","true");
ImportColumnFromCSV("Elevation","3","true");
ImportColumnFromCSV("Latitude","4","true");
ImportColumnFromCSV("Longitude","5","true");
SwitchToSourceTab(0);
join("CDEC - Event Data0","CDEC_stations.csv","Latitude");
join("CDEC - Event Data0","CDEC_stations.csv","Longitude");
PublishToWS("WINGS Portal","TEST_CDEC_WEATHER_82","CDEC - Event Data0");
KARMA Generates Data Processing Script

User demonstrates data processing for 1 day

KARMA Script works for any number of days
Publishing Processed Data to WINGS
Semantic Metadata for Input Files

```xml
<?xml version="1.0" encoding="UTF-8" ?>

<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:dc="http://www.isi.edu/dc/ontology.owl#"
    xmlns:dcdm="http://www.isi.edu/dc/Water/ontology.owl#"
    xmlns="http://www.isi.edu/dc/Water/library.owl#">

    <dcdm:Daily_Sensor_Data rdf:ID="FILENAME">
        <dcdm:forDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">DATE</dcdm:forDate>
        <dcdm:forSite rdf:datatype="http://www.w3.org/2001/XMLSchema#string">STATIONID</dcdm:forSite>
        <dcdm:siteLatitude rdf:datatype="http://www.w3.org/2001/XMLSchema#float">LATITUDE</dcdm:siteLatitude>
        <dcdm:siteLongitude rdf:datatype="http://www.w3.org/2001/XMLSchema#float">LONGITUDE</dcdm:siteLongitude>
        <dcdm:slope rdf:datatype="http://www.w3.org/2001/XMLSchema#float">SLOPE</dcdm:slope>
        <dcdm:depth rdf:datatype="http://www.w3.org/2001/XMLSchema#float">DEPTH</dcdm:depth>
        <dcdm:flow rdf:datatype="http://www.w3.org/2001/XMLSchema#float">FLOW</dcdm:flow>
    </dcdm:Daily_Sensor_Data>

</rdf:RDF>
```
Semantic Metadata for Input Files

<?xml version="1.0" encoding="UTF-8"?>

<rdf:RDF
    xmlns:dc="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:dc:sioc="http://www.w3.org/2007/06/sw-vocab-rdfs#"
    xmlns:dc:foaf="http://xmlns.com/foaf/0.1#"
    xmlns:sioct="http://rdfs.org/sioc/types#"
    xmlns:sioc:xt="http://rdfs.org/sioc/typesxt#"
    xmlns:sioc:rdf="http://rdfs.org/sioc/rdf#"
    xmlns:sioc:sioc="http://rdfs.org/sioc/rdf#"/>

<dc:Daily_Sensor_Data rdf:ID="FILENAME">
    <dc:forDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2003-10-27</dc:forDate>
    <dc:siteLatitude rdf:datatype="http://www.w3.org/2001/XMLSchema#float">LATITUDE</dc:siteLatitude>
    <dc:siteLongitude rdf:datatype="http://www.w3.org/2001/XMLSchema#float">LONGITUDE</dc:siteLongitude>
    <dc:slope rdf:datatype="http://www.w3.org/2001/XMLSchema#float">SLOPE</dc:slope>
    <dc:depth rdf:datatype="http://www.w3.org/2001/XMLSchema#float">DEPTH</dc:depth>
    <dc:flow rdf:datatype="http://www.w3.org/2001/XMLSchema#float">FLOW</dc:flow>
    <dc:barpress rdf:datatype="http://www.w3.org/2001/XMLSchema#float">760</dc:barpress>
</dc:Daily_Sensor_Data>
</rdf:RDF>
Workflows with WINGS

[Gil et al JETAI’11; Gil et al IEEE-IS’11; Gil et al e-Science’09; Kim et al JCC’08]
WINGS Received Metadata from KARMA
WINGS Received Metadata from KARMA

Metadata automatically associated with each input file
Workflow

1. Integrated Hourly Data (by Day)
2. Owens-Gibbs Model
3. O’Connor-Dobbins Model
4. Churchill Model
5. Compute Reaeration Parameters
6. Reaeration Rate (by Day)
7. Compute Metabolism
8. Net Daily Metabolism
Using Metadata in Workflow Execution

```xml
<dom:Hydrolab_Sensor_Data rdf:ID="Hydrolab-CDEC-04272011">  
  <dom:siteLongitude rdf:datatype="float">-120.931</dom:siteLongitude>  
  <dom:siteLatitude rdf:datatype="float">37.371</dom:siteLatitude>  
  <dom:dateStart rdf:datatype="date">2011-04-27</dom:dateStart>  
  <dom:forSite rdf:datatype="string">MST</dom:forSite>  
  <dom:numberOfDayNights rdf:datatype="int">1</dom:numberOfDayNights>  
  <dom:avgDepth rdf:datatype="float">4.523957</dom:avgDepth>  
  <dom:avgFlow rdf:datatype="float">2399</dom:avgFlow>  
</dom:Hydrolab_Sensor_Data>
```
Using Metadata in Workflow Execution

```xml
<dcdom:Hydrolab_Sensor_Data rdf:ID="Hydrolab-CDEC-04272011">
  <dcdom:siteLongitude rdf:datatype="float">-120.931</dcdom:siteLongitude>
  <dcdom:siteLatitude rdf:datatype="float">37.371</dcdom:siteLatitude>
  <dcdom:dateStart rdf:datatype="date">2011-04-27</dcdom:dateStart>
  <dcdom:forSite rdf:datatype="string">MST</dcdom:forSite>
  <dcdom:numberOfDayNights rdf:datatype="int">1</dcdom:numberOfDayNights>
  <dcdom:avgDepth rdf:datatype="float">4.523957</dcdom:avgDepth>
  <dcdom:avgFlow rdf:datatype="float">2399</dcdom:avgFlow>
</dcdom:Hydrolab_Sensor_Data>
```
Using Metadata in Workflow Execution

Metadata

<dom:Hydrolab_Sensor_Data rdf:ID="Hydrolab-CDEC-04272011">
  <dom:siteLong rdf:datatype="float">-120.931</dom:siteLongitude>
  <dom:siteLatitude rdf:datatype="float">37.371</dom:siteLatitude>
  <dom:dateStart rdf:datatype="date">2011-04-27</dom:dateStart>
  <dom:forSite rdf:datatype="string">MST</dom:forSite>
  <dom:numberOfDayNights rdf:datatype="int">1</dom:numberOfDayNights>
  <dom:avgDepth rdf:datatype="float">4.523957</dom:avgDepth>
  <dom:avgFlow rdf:datatype="float">2399</dom:avgFlow>
</dom:Hydrolab_Sensor_Data>

Setting Parameters

Choosing Models

O'Connor-Dobnig Model
Churchill Model
Reaeration Empirical
Reaeration Param
Metabolism Calc Empirical
Metabolism EDM
CreatePlots
CR24
GPP
NHH
PIK
Photo_Res
Sum_CorDO
### Workflow Results

Here is a screenshot of a workflow result page showing the details of a run.

#### My Runs

- **Template:** AquaFlow-1
- **Run ID:** 64
- **Project ID:** AquaFlow-1_Run_4770628

#### Result Browser

- **Data:** Input (4)
  - Hydrolab_Data
  - InputDepthFile
  - Latitude
  - Longitude

- **Data:** Intermediate (7)
  - MeanNightDO
  - OutputDailyParams
  - OutputHourlyAveragedData

#### Progress

- **Start Time:** 9:35:50 am, Mar 11, 2011
- **End Time:** 9:07:11 am, Mar 11, 2011
- **Status:** Finished!
Workflow Results Have Metadata

WINGS automatically generates metadata for each output file

```xml
<dcdom:Metabolism_Results rdf:ID="Metabolism_Results-CDEC-04272011">
  <dcdom:siteLong rdf:datatype="float">-120.931</dcdom:siteLong>
  <dcdom:siteLatitude rdf:datatype="float">37.371</dcdom:siteLatitude>
  <dcdom:dateStart rdf:datatype="date">2011-04-27</dcdom:dateStart>
  <dcdom:forSite rdf:datatype="string">MST</dcdom:forSite>
  <dcdom:numberOfDayNights rdf:datatype="int">1</dcdom:numberOfDayNights>
  <dcdom:avgDepth rdf:datatype="float">4.523957</dcdom:avgDepth>
  <dcdom:avgFlow rdf:datatype="float">2399</dcdom:avgFlow>
</dcdom:Metabolism_Results>
```
SELECT ?url WHERE {
?data dcdom:usedAlgorithm dcdom:ODM .
?data rdf:type dcdom:Metabolism_Estimates .
?data wflow:hasLocation ?url
}
Aquatic Photosynthesis

Models of gross primary production (GPP), community respiration (CR24)

Sensors
Aquatic Photosynthesis

Models of gross primary production (GPP), community respiration (CR24)

Sensors

Workflow Results
Aquatic Photosynthesis

Models of gross primary production (GPP), community respiration (CR24)
Summary

- Tools for end-users
- End to end support
- Data import, cleaning, integration
- Automated workflow execution
- Captures metadata provenance
Related Work

- Data integration:
  Data Wrangler [Kandel et al 2011]
  Google Refine [Huynh et al]

- Workflow systems:
  VisTrails [Howe et al 2008],
  Kepler [Barseghian et al 2010]

- Many tools generate provenance metadata, often in RDF
  - None generate other kinds of metadata
  - None use metadata to configure models
California’s Central Valley Water Project

Complex network of rivers, reservoirs, canals, groundwater basins

Grand/Long-range vision
–Observe, model, manage water resources to optimize stream ecology while sustaining society’s water needs
California Data Exchange Center
<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>TempF</th>
</tr>
</thead>
<tbody>
<tr>
<td>20091020</td>
<td>0</td>
<td>65.2</td>
</tr>
<tr>
<td>20091020</td>
<td>100</td>
<td>64.9</td>
</tr>
<tr>
<td>20091020</td>
<td>200</td>
<td>64.5</td>
</tr>
<tr>
<td>20091020</td>
<td>300</td>
<td>64.2</td>
</tr>
<tr>
<td>20091020</td>
<td>400</td>
<td>63.8</td>
</tr>
<tr>
<td>20091020</td>
<td>500</td>
<td>63.4</td>
</tr>
<tr>
<td>20091020</td>
<td>600</td>
<td>63.1</td>
</tr>
<tr>
<td>20091020</td>
<td>700</td>
<td>62.8</td>
</tr>
<tr>
<td>20091020</td>
<td>800</td>
<td>62.7</td>
</tr>
<tr>
<td>20091020</td>
<td>900</td>
<td>62.8</td>
</tr>
<tr>
<td>20091020</td>
<td>1000</td>
<td>63.2</td>
</tr>
<tr>
<td>20091020</td>
<td>1100</td>
<td>63.9</td>
</tr>
<tr>
<td>20091020</td>
<td>1200</td>
<td>64.6</td>
</tr>
</tbody>
</table>