Formally Specifying and Validating Data Assurance Properties for Real-Time Data Processing Applications
Irbi Gallego1, Santonu Goswani1
Mentors: Ann Gates2, Craig Tweedie2
Center of Excellence for Sharing Resources for the Advancement of Research and Education through Cyberinfrastructure
The University of Texas at El Paso, El Paso TX, 79968, USA
Emails: [irbi.g, goswani2]@miners.utep.edu; [agates, ctweedie]@utep.edu

Abstract
As global environmental conditions rapidly change, environmental scientists are using more data-intensive field-based instrument technologies to understand such changes. As the environmental sciences become more data driven, a need to develop mechanisms and procedures to verify the integrity of the data has increased. The quality of the data integrity verification mechanisms is as good as the quality of the properties specified. This research is investigating an approach to support scientists and non-software engineering domain experts in their efforts to specify "formally" data assurance properties. The work reported here describes an approach to formally specify real-time data properties and provides visualization of data properties that could allow users to validate whether the specified property captures the desired intent.

Visualization of a Formally Specified Property
- Once a practitioner specifies a property, he/she will be given an option to visualize the property to determine if the property captures the intended meaning.
- The proposed approach introduces a visualization technique based on traces to visualize time and data properties.

Figure 1: Prototype visual representation of a data property

| Reference values between wavelength value w<600 and w>1050 must always be positive |

Scientific Application
- For environmental scientists performing field-based research in the Arctic, it is difficult to identify corrupted data at real time.
- Corrupted data can result from adverse weather conditions, faulty equipment, or human error.
- In a common scenario:
  - The environmental scientist spends a day obtaining field data, with no indication about the quality of the data.
  - Once the data are obtained, the scientist spends another day analyzing and processing the data in the laboratory.
  - As a result of the analysis, the scientist may only then determine that the obtained data are corrupted.

The scientist must collect a new set of data and, as a result, loses time.

The proposed approach can help prevent such scenarios from occurring.

Motivation
- Environmental scientists perform experiments at hard-to-reach remote sites in order to understand environmental issues regarding climate change.
- The amount of data acquired at real-time by field-based instrument technologies has increased.
- It is vital to provide a mechanism to assure at real time that the data being obtained by the field-based instrumentation are correct.

The quality of any real-time data assurance mechanism is as good as the quality of the properties specified. Mechanisms that support verification and validation of properties are essential.

Research Goals
- The goals of this research are:
  - Define an approach to formally specify real-time data assurance properties.
  - Provide a representation of data assurance properties that would allow users to validate whether the specified property captures the desired intent.
  - The proposed work is answering the following research questions:
    - What real-time data properties are relevant to environmental science sensor data and which rely on temporal constraints and other knowledge?
    - What types of behavior patterns can be associated with the categories of real-time properties?
    - How can patterns facilitate data properties elicitation?
    - What is the mapping of patterns, if applicable, to a formal specification?
    - How can formally specified scientific real-time data properties be represented for validation purposes?

Formally Specifying Data Properties
Findings from a case study conducted at Barrow-Arctic Science Consortium (BASC) in June 2008 identified the need to specify three types of data assurance properties:
- **Time Properties**: properties used to specify timing constraints about entities.

- **Data Properties**: properties used to specify characteristics about the data value themselves.
  - A data assurance (DA)-SPS and a data assurance (DA)-CP will be developed for specification of data properties.
  - The original SPS and CP have to be redefined to account for properties about data attributes.

- **Instrumentation Properties**: properties used to specify the expected behavior for the data gathering instruments.
  - An instrumentation property can be either a data property or a timing property depending on the entity over which the behavior is being specified.
  - Instrumentation properties will be specified using a combination of SPS, DA-SPS, CP, and DA-CP.

Case Study
- A case study was performed in collaboration with the System Ecology Laboratory at The University of Texas at El Paso in the Summer 2008 at the Arctic research site in Barrow, Alaska.
- The purpose of the case study was to document the types of real-time data assurance properties that are of interest to environmental scientists and investigate how these properties can be captured.
  - **The case study**: identified three types of properties: data properties, timing properties, and instrumentation properties.
  - The results obtained from the initial stage of the case study were used to develop a data assurance real-time monitoring tool to verify formally specified DA-SPS data properties.
  - The real-time data monitoring tool was incorporated into the train system to evaluate the feasibility of the proposed approach.
    - The tool correctly quality-tagged at real-time the data as it was being streamed from the train system into the base station.
    - The tool also instantly verified the quality of the data that was being transferred in real-time from the train system into the base station.
    - The tool identified inconsistencies in the data quality data.

Findings were used to improve data current models.

References

This work is supported by National Science Foundation award # CNS-0540592 and award # HRL-0734823.