JRA3: New Mining and Integration Tools
Cycle One Update

Dugan Witherick

University College London

VAMDC Cycle One Project Meeting @ OU

Purpose of JRA3

- Objectives
  - To complement JRA1 and JRA2 by building the query protocols that will access those published AM data and then to design software that will handle and process those data.
- Description of Work
  - Through the activities of JRA1 and JRA2, the AM resources will be searchable and will provide information in a standardised way. The following step is to build the query protocols that will access those published AM data and then to design software that will handle and process those data.

Task 1: Registry Queries (Summary)

- Led by CNRS
- To implement the protocols to query the registry at a fine level of granularity
  - For example by selecting resources by they content through keyword search
- Not due to start until Cycle Two

Task 2: Tools for the Manipulation of Data (Summary)

- Led by Koln
- Overall Plan of Work
  - Queries return data organised by schemas defined in JRA1
  - Complex handling of the XML requiring Specific Tools
  - Two types of tools identified:
    - Cross-matching of data
    - Cross-federation
  - Will be made available to download in SA1 to end users and developers
  - Support to adapt those tools to specific applications will be provided in SA2.
  - We plan to provide libraries to allow users to develop their own applications

Task 2: Tools for the Manipulation of Data (Cycle One Update)

- Task 2.1: Make Prototype of Data cross-identification software based on current XSAMS schema (link with JRA1/JRA2)
  - Initial Test for Cycle 1 will be between BASECOL and CDMS databases.
- Paris JRA1/2 (18-19 Feb)
- Work on cross identification tool by Ljerka Nenadovic (between BASECOL and CDMS, and using XSAMS)

Task 3: Advanced data mining services (Summary)

- Led by UCL
- Task will investigate optimal strategies to best mine these AM data resources to both advance the creation of new AM fundamental data, and by providing stream lined automated access to appropriate AM data targeted at specific user groups.
- Would investigate the provision of application services wrapping complex work flows combining AM data access, manipulation, and integration into user processing chains
  - e.g. in solar physics, astro-biology/ chemistry and so forth.
- Linked into to JRA1 and JRA2

Task 3: Advanced data mining services (Cycle One Update)

- Development of advanced data mining tools is focusing on the deployment of an XSAMS compliant HITRAN database
- Significant efforts made by Christian Hill (UCL)
  - Input into the XSAMS schema
  - Attendance of Paris JRA1/2 meeting (18-19th February, 2010)
  - Attendance of Uppsala workshop for developing software tools related to the VAMDC project (6th-9th April, 2010)

Task 3: Advanced data mining services (Cycle One Update cont.)

- http://voparis-twiki.obspm.fr/twiki/bin/view/VAMDC/HitranProgress
- Extensions to XSAMS for HITRAN
- Experiments with a new way of representing Molecular States in XSAMS
- Development of pyHAWKS program to convert HITRAN data to XSAMS-compliant XML
- Further development and maintenance of the XSAMS2HITRAN program which does the conversion the other way
  - this will be the most useful from the VAMDC-user point of view since the output is in the native 160-character line format understood by existing code

Task 3: Advanced data mining services (Cycle One Update cont.)

- A prototype version of the HITRAN database, represented in a MYSQL table is being optimized.
  - Can be queried by molecule type, isotopologue, wavenumber range, and line-strength.
- Looking to migrate the HITRAN database to a queryable MySQL database
  - Hosted at UCL
  - Return XSAMS-compliant output for limited testing on a range of molecules.
  - Prototype service but should eventually be deployed using tools developed in JRA2

Task 3: Advanced data mining services (Cycle One Update cont.)

- Development of science use cases for e-HITRAN workflows
  - Identifying users/data miners and creating science use cases based on their current and planned future use
  - Have begun a dialogue with MSSL (Kevin Benson) on technical requirements of workflows (python/taverna)
- Jonathan Tennyson and Christian Hill to be attending 11th Bienniel HITRAN Conference (16-18 June 2010)
  - Information for science use cases will be collated at this meeting and digested shortly afterwards