VAMDC

Virtual Atomic and Molecular Data Centre

D8.2

–

Mining/Integration Report 1

Version 0.3

Grant agreement no: 239108

Combination of Collaborative Projects & Coordination and Support Actions

Project Information

Project acronym: VAMDC

Project full title: Virtual Atomic and Molecular Data Centre

Grant agreement no.: 239108

Funding scheme: Combination of Collaborative Projects & Coordination and Support Actions

Project start date: 01/07/2009

Project duration: 42 months

Call topic: INFRA-2008-1.2.2 Scientific Data Infrastructure

Project web sites: http://www.vamdc.eu

http://voparis-twiki.obspm.fr/twiki/bin/view/VAMDC/WebHome

Consortium:

Beneficiary Number *	Beneficiary name	Beneficiary short name	Country	Date enter project**	Date exit project**
1(coordinator)	Centre National de la Recherche Scientifique	CNRS	France	Month 1	Month 42
2	The Chancellor, Masters and Scholars of the University of Cambridge	CMSUC	UK	Month 1	Month 42
3	University College London	UCL	UK	Month 1	Month 42
4	Open University	OU	UK	Month 1	Month 42
5	Universitaet Wien	UNIVIE	Austria	Month 1	Month 42
6	Uppsala Universitet	UU	Sweden	Month 1	Month 42
7	Universitaet zu Koeln	KOLN	Germany	Month 1	Month 42
8	Istituto Nazionale di Astrofisica	INAF	Italy	Month 1	Month 42
9	Queen's University Belfast	QUB	UK	Month 1	Month 42
10	Astronomska opservatorija	AOB	Serbia	Month 1	Month 42
11	Institute for Spectroscopy RAS	ISRAN	Russian Federation	Month 1	Month 42
12	Russian Federal Nuclear Centre All-Russian Institute of Technical Physics	RFNC-VNIITF	Russian Federation	Month 1	Month 42
13	Institute of Atmospheric Optics	IAO	Russian Federation	Month 1	Month 42
14	Corporacion Parque Tecnologico de Merida	CPTM	Venezuela	Month 1	Month 42
15	Institute of Astronomy of the Russian Academy of Sciences	INASAN	Russian Federation	Month 1	Month 42

This project is funded under *"Combination of Collaborative Projects and Coordination and Support Actions"* Funding Scheme of The Seventh Framework Program of the European Union

Document

Deliverable number:	D8.1
Deliverable title:	Mining/Integration Plan
Due date of deliverable:	September 2009
Actual submission date:	8th September 2010
Authors:	J. Tennyson, D.J. Witherick, M.L. Dubernet
Work Package no.:	WP8-JRA3
Work Package title:	New Mining and Integration Tools
Work Package leader:	UCL
Lead beneficiary:	UCL
Dissemination level:	PU
Nature:	Report
No of pages (incl. cover):	

Abstract

The objective of D8.1 is to describe VAMDC New Mining and Integration Tools Plan on PM3. This plan corresponds to Activities in WP8: JRA3 "New Mining and Integration Tools". This plan is included in the VAMDC Project Plan.

Versioning and Contribution history

Version	Date	Reason for modification	Modified by
V0.1	**20/07/2010**	**Section 5**	**D. J. Witherick, L. Nenadovic**
V0.1	**19/08/2010**	**Draft D8.2**	**M.L. Dubernet**
V0.2	**04/09/2010**	**Modification of Section 5 to better reflect WP8-T3 activities**	**M.L. Dubernet**
V0.3	**04/09/2010**	**New modfications in Task 3 activities, reports**	**D.J. Witherick**

Final Version (vx.x) released by		Circulated to	
Name	Date	Recipient	Date
M.L. Dubernet	8th September 2010	Mrs Asero	8th September 2010

WP8 ACTIVITIES DESCRIPTION

Work package number	8		Start date or starting event:		3						
Work package title	JRA3: New mining and Integration Tools										
Activity Type	RTD										
Participant id	1	3	7	12							
Person-months per beneficiary: (Total = EU + Node Contributions)	12	36	18	6							

Table of Content

1. WP8 Objectives

This JRA will develop extensions to the baseline infrastructure. Key objectives are the design of advanced data mining tools and the design of cross-matching and cross-federating tools, providing sophisticated integrated science services aimed at maximising the scientific utility to the end user community of the VAMDC services.

WP8 Leader is UCL(3)

2. WP8 Milestones and Deliverables

Milestones

M8.1	Technical Meetings	WP8	UU	Months 5,10, 16, 22, 28, 34, 40, 42	Minutes. Presentations on internal Website
M8.2	Evaluation of softwares	WP8	UU	Months 10, 22, 34	

Deliverables

D8.1 Mining and Integration Tools Plan (PM 3)
D8.2 Mining and Integration Tools Report to be included in report to the EU – Year 1 (PM 10)
D8.3 Mining and Integration Tools Report to be included in report to the EU – Year 2 (PM 22)
D8.4 Mining and Integration Tools Report to be included in report to the EU – Year 3 (PM 34)

D8.5 Final Report of Mining and Integration Tools to be included in final report to the commission (PM41)
Annual Mining & Integration Plan revisions included in Revised Annual VAMDC Project Plans – Year 1,2,3

3. WP8 Tasks Description

WP8 Leader (co)		
Task Number	Leader	Other Partners
1	M. Doronin (CNRS: LPMAA)	RFNC-VNIITF
2	S. Schlemmer (KOLN)	CNRS:LPMAA
3	J. Tennyson (UCL)	UCL/MSSL

Description of work (possibly broken down into tasks)

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.

Task1: Registry Queries (lead by CNRS(1) with (12))
We will need to use protocols to query the registries at a fine level of granularity. Indeed we don't wish to only find resources having implemented a type of service such as SSAP or TAP, but rather be able to select resources according to their content through key words. The purpose of Task 1 is to implement those protocols.

Task 2: Tools for Manipulation of Data (lead by KOLN(7) with (1))
Our queries will return data organised according to schemas defined in JRA1. Those schemas will be quite complex because they will reproduce all the scientific concept attached to the data. Therefore the handling of the XML files will be complex and will require specific tools. For now we identify too main generic tools: one performing cross-matching of data and one performing cross-federation of data. These tools are particularly difficult because they require to compare the content of many fields in the schema. Those generic tools will be made available for download in SA1 to the 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 3: VAMDC advanced data mining services (lead by UCL(3))
With the deployment of a vast range of high value data services through the standard VAMDC infrastructure, this 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 (for the astronomy community benefiting from the availability of high energy data from satellites such as Swift, XMM, Chandra, who require specific atomic data for high excitation species of elements such as iron). This task 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.

4. WP8 Tasks Description for Period 1

Full task activities are detailed at the VAMDC wiki on the WP8/JRA3 pages – see http://voparis-twiki.obspm.fr/twiki/bin/view/VAMDC/WP8

Task 1: Registry Queries

- This task is not due to start until Cycle 2

Task 2: Tools for Manipulation of Data

- 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.

Task 3: VAMDC advanced data mining services

- Development of use cases for workflows based around e-HITRAN
  - Identify current and future users/data miners of the HITRAN database and its upcoming successor, e-HITRAN
  - Survey the current and future HITRAN/e-HITRAN users as to how they do/would mine the information contained in the database
  - Develop a series of use cases based upon the responses of users

- Development of technical requirements of workflows
  - Begin dialogue with MSSL on the technical requirements for workflows to support the use cases

5. WP8 Tasks Report for Period 1

VAMDC
Periodic Report Template (per Workpackage)

Period: 01/07/2009 – 30/06/2010
WorkPackage: WP8 New Mining and Integration Tools
WorkPackage Leader and co-Leader: Jonathan Tennyson (UCL), Dugan Witherick (UCL)
Participants in the WorkPackage: UCL, CNRS in Period 1 (no participation from KOELN, RFNC-VNIITF)

Part 1

A summary of progress towards objectives and details for each tasks

Task 1: Registry Queries (leader: CNRS)
The objective of this task is to implement the protocols necessary to query the registries to a fine level of granularity and it is not due to start until cycle 2. Nevertheless, discussions between partners in WP4 and WP6 led to the conclusion that International Virtual Observatory Alliance (IVOA) standards would be adopted meaning that in effect this task has now been completed.

Task 2: Tools for the Manipulation of Data (leader: CNRS)
The objective of this task is to develop tools for the purpose of cross-matching and cross-federation of data stored in the format defined in WP6. A prototype tool to cross-match data stored in the XML Schema for Atoms, Molecules and Solids (XSAMS) format has been developed and can successfully match energy levels between data (e.g. BASECOL, CDMS) representing the same species based on quantum number values. XSAMS has continued to evolve through cycle one (WP6 Task 1) and this prototype program has continued to be extended to support each new version of the schema.

Task 3: VAMDC Advanced Data Mining Services (leader: UCL)
The objective of this task is to develop advanced data mining services for VAMDC, to enable specific user groups streamlined automated access to appropriate AM data. The development of these services relies on the deployment of the basic infrastructure of VAMDC (registry, data access services); widespread deployment of the basic infrastructure is expected during period 2 and so the widespread deployment of these data mining services is deferred until period 3. The development of these services also requires detailed use cases to be produced for specific user groups. Progress towards the development of these use cases has focused on the High-resolution Transmission Molecular Absorption (HITRAN) users, given that it is a relatively large community, with a wide variety of uses. In the last month of cycle one, at the 11th Biennial HITRAN Conference, a survey was conducted on the way that the HITRAN user community queries the database and how they currently use and would like to use the data in the future. This survey data will be used to form the basis of data mining use cases, which will be developed in detail in period 2.

Significant results (Activities and Deliverables)

Task 1:
The adoption of IVOA standards means that this task has been effectively completed.

Task 2:
Initially the XSAMS v 0.1 schema was considered. The experiences with this schema led to a proposal of a less nested schema for the description of molecular states. Following the XSAMS meeting in Japan, where a modified schema was discussed, methods for extraction of information from this model were developed. C. Hill (UCL) proposed an alternative case-by-case approach and the prototype was further extended to include manipulation of data in this schema.

Implementing a prototype for the various versions of the data model has been important in ensuring that the data model that will ultimately be adopted as the standard for VAMDC contains all aspects necessary for cross-identification.

Internal Deliverables

a) The development of a prototype Java based tool for querying BASECOL database and returning the results in XSAMS format.
b) Extension of the prototype application to enable cross-matching of energy levels based on quantum number values from data stored in the XSAMS format.

Task 3:
A survey has been performed on the HITRAN user community at the 11th Biennial HITRAN conference and has resulted in the identification of a several generic use cases for AM data. This survey data will form the basis of use cases, which will be developed in detail in period 2.

Deviations from the contract (Annex I) and reasons for them (if applicable)

There is deviation from contract related mainly to task 3 because no critical objectives have been achieved in Task 3 (see below)–
There is also a deviation from contract with respect to MP from KOELN, RFNC-VNIITF, UCL who have shifted MP from WP8 to WP7, WP6 because data need to be characterised and published before being handled, queried or registered. This has no real consequence on the development of tools in Task 2. As task 1 is completed, lack of contribution of RFNC-VNIITF has no influence. Shift of MP-UCL from WP8-Task 3 was a necessity (see explanations below) with respect to deployment of HITRAN. Shift of MP-KOELN is the same because of deployment of CDMS.

Failures to achieve critical objectives and/or not being on schedule and reasons for them (if applicable)

Task 3: VAMDC Advanced Data Mining

a) The development of advanced data mining services relies on the deployment of the basic infrastructure of VAMDC (registry, data-access services), which is expected during period 2. Manpower assigned to this task has instead been shifted to WP6 Task 1 (Data Models and Schemas) and WP7 Tasks 1 and 3 (Publishing Tools).
b) The user survey from the HITRAN Biennial conference has yet to be fully analysed since the conference took place in the last month of period 1 but provided the best forum for collecting the required data.

Proposed corrective actions (if applicable)

Task 3: VAMDC Advanced Data Mining

a) Detailed use cases for data mining services will be developed in period 2 and one or more of these cases will be implemented by WP4 to establish the techniques. Widespread deployment of data mining services is deferred to period 3.
b) Analysis of the HITRAN user survey will be completed in period 2.

(approximate length of Part 1: 2 pages)