VAMDC

Virtual Atomic and Molecular Data Centre

D8.3

–

Mining/Integration Report 2

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	CTPM	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:	D8.3
Deliverable title:	Mining/Integration Report 2
Due date of deliverable:	June 2011
Actual submission date:	August 2011
Authors:	D. Witherick, J. Tennyson, L. Nenadovic, B. Schmitt, K. Benson, F. Kosmala, M.L. Dubernet and WP8 team
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 D8.3 is to describe VAMDC Science/Technical Report for Cycle 2. This report corresponds to Activities in WP8: JRA3 "New Mining and Integration Tools". This report will be included in the VAMDC Periodic Report for Cycle 2.

Versioning and Contribution history

Version	Date	Reason for modification	Modified by
V0.1	**June 2011**	**Compilation of Nodes Contributions to WP8**	**F. Kosmala**
V0.1	**July 2011**	**WP8 report for P2**	**D. Witherick**
V0.1	**July/August 2011**	**Making of D8.3 Document – Inclusion of above P2 report**	**F. Kosmala**
V0.2	**August 2011**	**Check about Content**	**M.L. Dubernet**
V0.3	**August 2011**	**More information from Grenoble on T2.2**	**B. Schmitt, D. Witherick**
V0.3	**August 2011**	**Final Check**	**M.L. Dubernet**

Final Version (v0.3) released by		Circulated to	
Name	Date	Recipient	Date
M.L. Dubernet	24th August 2011	Mrs Asero	24th August 2011

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

WP8 activities description 5
Table of Content 5
1. WP8 Objectives 5
2. WP8 Milestones and Deliverables 5
3. WP8 Tasks Description 6
4. WP8 Tasks Plans for Period 2 7
5. WP8 Tasks Reports for Period 2 8

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 Plans for Period 2

Period: 01/07/2010 – 30/06/2011
WorkPackage: WP8 New Mining and Integration Tools
WorkPackage Leader and co-Leader: Jonathan Tennyson (UCL), Dugan Witherick (UCL)
Participants in the WorkPackage: UCL, CNRS, KOELN

Part 1

Objectives and details for each task in Year 2.

Task 1: Registry Queries – task CLOSED – IVOA Standards

Task 2: Tools for the Manipulation of Data (leader: CNRS)

T2.1 Improvement of Prototype of Data cross-identification software based on current XSAMS schema (link with JRA1/JRA2) handling more type of data (extending further than energy levels comparison) and Connection to Astrophysical Users - Inclusion of Tool in user applications –

T2.2 Define specifications both scientific and technical for a software tool allowing to handle both gas spectroscopy and solid spectroscopy data - Connection to Astrophysical and Planetology Users

Task 3: VAMDC Advanced Data Mining Services (leader: UCL)

The objectives of Task 3 of WP8, the development of VAMDC advanced data mining services, rely on the deployment of the basic infrastructure of VAMDC (registry, data-access services). Wide-scale deployment of the data-mining services is deferred to Period 3.
In Period 2, we shall develop in detail the use cases (query of gas phase data and/or solid spectroscopy data) to be implemented in Period 3. One or more of these cases will
be implemented by WP4 (task 2.2 in that WP) during period 2 to establish
the techniques.

5. WP8 Tasks Reports for Period 2

Period: 01/07/2010 – 30/06/2011
WorkPackage: WP8 New Mining and Integration Tools
WorkPackage Leader and co-Leader: J. Tennyson (UCL), D. Witherick (UCL)
Participants in the WorkPackage: UCL, CNRS (no participation from KOELN, RFNC-VNIITF)

Part 1

A summary of progress towards objectives and details for each tasks

Task 1: Registry Queries
The objective of this task was to implement the protocols necessary to query the registries to a fine level of granularity. This work was due to commence in cycle two; however the adoption of the International Virtual Observatory Alliance (IVOA) standards in cycle one meant that the task was effectively completed.

Task 2: Tools for the Manipulation of Data
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.

During cycle one, a prototype tool was developed to cross-match/cross-federate XSAMS formatted data from CDMS (spectroscopic data) and BASECOL (collisional data); in cycle two, this tool was developed further and what follows is a summary of some of the new additions/changes:

- Significant improvements were made in the design and robustness of the tool, as well as the GUI, to make it more reliable and easier to use.
- The tool was updated to follow VAMDC-XSAMS as well as gaining registry and database querying features.
- The XSAMS manipulation tool gained import and graphing functionality.
- The tool has gained the ability to reduce XSAMS down to all data related to a specific molecule.

This tool has now left the development phase and has been made available to the wider VAMDC community. In addition, some of the general purpose XSAMS manipulation functions are in the process of being packaged up as libraries for release during cycle three.

During cycle two, progress was also made towards extending the Grenoble Astrophysics and Planetology Solid Spectroscopy and Thermodynamics (GhoSST) database service to support interaction with the VAMDC infrastructure. GhoSST was extended to support querying using XSAMS keywords which required the existing interface to be extended to cross-match between XSAMS and SSDM keywords; some of these translations were simple whilst others required indirect computation. A VAMDC output engine was also developed, enabling queries to the service to generate VAMDC-XSAMS compatible data. The GhoSST VO service has been registered in the VAMDC developmental registry and is able to accept simple queries directly at http://ghosst-prod.obs.ujf-grenoble.fr/vamdc/tap/sync

Task 3: VAMDC Advanced Data Mining Services
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 cycle two plan for this task was to develop the detailed use cases (i.e. the query of gas phase data and/or solid spectroscopy data), which would then be implemented during cycle three as workflows. However the development of the use cases and the subsequent success of the solutions would be dependent upon the mechanism used to enact the workflow and ultimately on the capabilities of the underlying VAMDC infrastructure. To this end, effort was focused on the development of the "workflow engine", consisting of a plugin module to the Taverna workflow management system, enabling certain VAMDC services (such as CEA wrapped applications and data services) to be included in Taverna workflows. By the end of cycle two, development had completed on a fully working prototype plugin which was capable of running simple workflows demonstrating the ability to interact with the VAMDC infrastructure. Further details of this prototype plugin and instructions on its use may be found at http://vamdc.mssl.ucl.ac.uk/taverna/vamdc/

Significant results (Activities and Deliverables)

Task 1:

The task has been closed since cycle 1.

Task 2:

a) Significant further development of the cross-matching tool developed during cycle one, improving its usability and extending its functionality. This tool has been released to the VAMDC community (called SPECTCOL) and the CASSIS Team (developing software for analysis of molecular spectra) from Toulouse have confirmed that it meets their needs for the analysis of spectra from the interstellar medium.
b) The extension of the GhoSST database service to support VAMDC keywords and to output data in the VAMDC-XSAMS compatible format.

Task 3:

The development of a prototype VAMDC plugin to the Taverna workflow engine to enable VAMDC services to be included in Taverna workflows.

Deliverables to EU

D8.1 Mining and Integration Tools Plan- DONE –
See http://www.vamdc.eu/public-deliverables/17-deliverables-wp8

D8.2 Mining and Integration Tools Report to be included in report to the EU – Year 1 – DONE –
See http://www.vamdc.eu/public-deliverables/17-deliverables-wp8

D8.3 Mining and Integration Tools Report to be included in report to the EU – Year 2 – DONE –
See http://www.vamdc.eu/public-deliverables/17-deliverables-wp8

Annual Mining & Integration Plan revisions included in Revised Annual VAMDC

Project Plans – Year 1,2 *See D1.2 and D1.5* **http://www.vamdc.eu/public-deliverables/12-deliverables-wp1**
Internal Deliverables All softwares (SPECTCOL Tool and Taverna Plugin) can be found either on the www.vamdc.eu/softwares or on http://voparis-twiki.obspm.fr/twiki/bin/view/VAMDC/Na1T1Tools

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

There have been no deviations from the contract.

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

Task 3: VAMDC Advanced Data Mining The cycle two plan for this task was to develop the detailed use cases to be implemented as workflows in cycle three. However it was deemed that knowledge of the eventual capabilities of the VAMDC "workflow engine" (the Taverna plugin) would act as initial selection criteria for use cases. Task three man power therefore focused on the development of the prototype Taverna plugin, the development of which was itself dependent upon the continual deployment of the VAMDC Infrastructure during cycle two.

Proposed corrective actions (if applicable)

Task 3: VAMDC Advanced Data Mining The development of the prototype Taverna plugin was completed towards the end of cycle two; and will be released to the limited VAMDC community during the early parts of cycle three. Feedback from these early users will provide us with an understanding of the limitations and capabilities of the workflows, as well as an idea of the type of workflows that end users will ultimately find useful. We will use this information to formulate the use cases and subsequently develop work flows which will be within the capabilities of the VAMDC workflow engine.