VAMDC

Virtual Atomic and Molecular Data Centre

# D8.5

**–**

# Final Mining/Integration Report

Version 1.0

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 |

VAMDC
Virtual Atomic and Molecular Data Centre

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

**<u>Document</u>**

| | |
|---|---|
| Deliverable number: | D8.5 |
| Deliverable title: | Final Mining/Integration Report |
| Due date of deliverable: | December 2012 |
| Actual submission date: | January 2013 |
| Authors: | D. Witherick, J. Tennyson, ML 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 D8.5 is to describe VAMDC Science/Technical activities for the whole project. This report corresponds to Activities in WP8: JRA3 "New Mining and Integration Tools".

**Versioning and Contribution history**

| **Version** | **Date** | **Reason for modification** | **Modified by** |
|---|---|---|---|
| **V0.1** | **Dec 2012** | **Preparation of DOcument** | **M.L. Dubernet** |
| **V0.1** | **Dec 2012** | **WP8 report** | **D. Witherick** |
| **V1.0** | **January 2013** | **D8.5 Document –** | **ML. Dubernet** |

| Final Version (v1.0) released by | | Circulated to | |
|---|---|---|---|
| Name | Date | Recipient | Date |
| M.L. Dubernet | 12th February 2012 | Mr Bodo | 12th February 2012 |
| | | | |

**Disclaimer:** The information in this document is subject to change without notice. Company or product names mentioned in this document may be trademarks or registered trademarks of their respective companies.

**All rights reserved**:

The document is proprietary of the VAMDC consortium members. No copying or distributing, in any form or by any means, is allowed without the prior written agreement of the owner of the property rights.

This document reflects only the authors' view. The European Community is not liable for any use that may be made of the information contained herein.

# 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:**<br>**(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<br>5, 10, 16, 22, 28, 34, 40, 42 | Minutes. Presentations on internal Website |
| M8.2 | Evaluation of softwares | WP8 | UU | Months<br>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 Final Tasks Reports

**Period**: 01/07/2009 – 31/12/2012
**WorkPackage:** (8) JRA3: New mining and Integration Tools
**WorkPackage Leader and co-Leader**: Jonathan Tennyson and Dugan Witherick
Participants in the WorkPackage: UCL (lead), CNRS/LPMAA, IPAG (Grenoble), KOLN

Part 1

A summary of progress towards objectives and details for each tasks

### Summary

The key objective of this work package has been to build extensions to the base VAMDC infrastructure to maximise the scientific utility to the end user community of the VAMDC services. The work package has employed a number of strategies to meet this objective ranging from building tools to manipulate the XSAMS data, enhancing web interfaces for specific scientific communities to utilise VAMDC infrastructure and, enabling workflows to be built that utilise VAMDC services and services from other, similar projects. The nature of this work package means that the tasks have been entirely dependent upon the decisions and work effort committed in work packages 6 and 7.

What follows is a summary of the work completed in each of the defined tasks of WP8.

### Task 1: Registry Queries

For the VAMDC services to be of use to the end user community, they must be able to query the registries at a fine level of granularity such as, for example, being able to select resources through keyword searches. The purpose of this task was to implement the protocols that are required to make this possible.

During cycle 1, discussions between the partners in WP4 and WP6 led to the conclusion that IVOA (International Virtual Observatory Alliance) standards would be adopted for the VAMDC project. The net effect of this on this task was that the protocols necessary to search the registry at a fine level of granularity (along with the corresponding software) had already been developed and did not need to be developed separately for VAMDC. Thus, this task was effectively closed at the end of cycle 1, despite not being scheduled to start until cycle 2.

### Task 2: Tools for Manipulation of Data

Queries to VAMDC services produce data in the VAMDC-XSAMS (XML Schema for Atoms, Molecules and Solids) format, which is a relatively complex XML schema, developed so that the files may reproduce all of the scientific concepts attached to the data. This means that the files could be complex to work with, not least for those end-users with no experience with work with XML files. To mitigate some of these issues, the objective of this task was to develop tools that enable end-users to manipulate the resulting VAMDC-XSAMS files. This section will describe this work in more detail.

*SPECTCOL (CNRS/LPMAA)*

SPECTCOL is a tool for VAMDC end-users to use in order to cross-match or cross-federate data formatted in the VAMDC-XSAMS format. The tool has a graphical user interface which has been refined following feedback from users and, was developed in Java so is cross-platform.

The original prototype of SPECTCOL was only capable of working with two data sets (namely BASECOL and CDMS) but when released to the user community in cycle 2, it was capable of working on all VAMDC-XSAMS data. Over the course of the project the tool has undergone development work to ensure that it is capable of working with the most recent versions of the VAMDC-XSAMS schema and, to provide additional functionality:

- Directly querying the registry and databases.
- Data importing.
- Graphing.
- The ability to reduce the XSAMS data down to all the data related to a specific molecule.

The SPECTCOL tool was used extensively by a number of data providers to check the Quality Assurance of their XSAMS output and, the CASSIS team, who were developing software for the analysis of molecular spectra, confirmed that the tool was suitable for their needs. On the completion of the development work, the SPECTCOL tool became part of the WP4-Task 5 software release (http://www.vamdc.eu/software )

*GhoSST (IPAG)*

The GhoSST (Grenoble Astrophysics and Planetology Solid Spectroscopy and Thermodynamics) database service provides astrophysicists and planetologists tools to search, explore, visualize and export data from the GhoSST database. In cycle two, work began on adapting GhoSST to make it compatible with the VAMDC infrastructure and by the end of that cycle the service was accepting VAMDC keywords and output data in the VAMDC-XSAMS format. During cycle three, the development of the VAMDC compatibility continued and expanded to supporting the Solid Spectroscopy Data Model (SSDM) so that it became possible to retrieve spectra or band list data of the same molecule/ion in the same spectral range for both gas and molecular solid thus demonstrating a Solid-Gas data inter-comparison service.

The development work on the service completed at the beginning of cycle 4 and was released as a public service at the end of September 2012, accessible at the following location
http://ghosst.obs.ujf-grenoble.fr/

CDMS Tools (Koln)

The Cologne Database for Molecular Spectroscopy (CDMS) contains a catalogue of radio frequency and microwave to far-infrared spectral lines of atomic and molecular species that (may) occur in the interstellar or circumstellar medium or in planetary atmospheres. The CDMS data providers (Koln) developed tools to extract information from VAMDC-XSAMS documents for display or to generate files that are commonly used in astrochemistry. These tools have been added to the CDMS web portal, which had already been extended to provide access to the VAMDC network in a form that's idea for the special needs of the spectroscopic and astronomical communities.

The CDMS web portal, as well as the additional tools developed for manipulating the data may be found at http://cdms.ph1.uni-koeln.de/DjCDMSdev/cdms /.

**Task 3: VAMDC advanced data mining services**

The objective of this task was to develop advanced data mining services for VAMDC to enable specific user groups streamlined automated access to appropriate atomic and molecular data. The development of these services was dependent upon the deployment of the basic VAMDC infrastructure (i.e. the registry and data-access services) so, as UCL was lead for this task and, the HITRAN (High Resolution Transmission Molecular Absorption) data provider, the opportunity was taken to survey the HITRAN user community (at the 11th Bienniel HITRAN Conference) on how they use and would like to use the HITRAN data in the future.

From the requirements capture exercise in cycle 1 and discussion with other partners, the decision was made to develop a VAMDC plugin for the Taverna workflow engine. Taverna is an "open source and domain-independent Workflow Management System" that enables users to build complex workflows using a simple, cross-platform graphical user interface. An example of its capability would be the simultaneous querying of multiple data sources, sending the results of which to a service which could cross-match the data and then pass to a numerical code. Taverna is heavily used in the bioscience and bioinformatics domains and it continues to receive funding for its development.

In cycle 2, work began on the development of the prototype VAMDC plugin to the Taverna workflow engine, to enable VAMDC services to be queried directly from the workflow. By the end of the cycle, a fully working prototype plugin had been produced and this plugin was capable of running simple workflows, including the use of a limited number of VAMDC data sources. During cycle 3, development of the plugin continued until a production ready version was produced and released to the VAMDC end-user community for testing and feedback. In addition a number of workflows were developed and released to the 'myExperiment' workflow sharing web site; workflows hosted on the 'myExperiment' web site can be downloaded directly from Taverna and may be used as a template for creating new workflows.

In cycle 4 the plugin entered maintenance stage with bug and feedback requests being monitored through the standard VAMDC support ticketing system. Efforts were made to improve the user support for the plugin by a) creating videos demonstrating the installation and use and b) creating supporting documentation on the VAMDC wiki. Steps were taken to improve community knowledge of the plugin by, for example, given a presentation at the France OV Workflow Working Group meeting (Paris November 2012).

The VAMDC plugin to Taverna may be found at http://voparis-twiki.obspm.fr/twiki/bin/view/VAMDC/TavernaUserGuide (along with extensive documentation) and the VAMDC workflows on the myExperiment site may be found at http://www.myexperiment.org/search?filter=CATEGORY%28%22Workflow%22%29&query=vamdc or by performing a search for VAMDC within the Taverna interface.

Significant results (Activities and Deliverables)

**Internal Deliverables**

**Task 1**

- The adoption of IVOA standards meant that existing tools and protocols could be used for this task.

**Task 2**

- The release of the SPECTCOL cross-matching and cross-federating tool.

- The extension of the GhoSST service to support the VAMDC infrastructure.
- The release of tools to manipulate XSAMS data on the CDMS web site.

**Task 3**

The release of the VAMDC plugin to the Taverna workflow engine.

**Deliverables to EU**

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

***D8.4 Mining and Integration Tools Report to be included in report to the EU – Year 3 – DONE –***
***See http://www.vamdc-project.vamdc.eu/public-deliverables/17-deliverables-wp8***

***Annual Mining & Integration Plan revisions included in Revised Annual VAMDC Project Plans – Year 1,2,3***
***See D1.2, D1.5, D1.7*** **http://www.vamdc-project.vamdc.eu/public-deliverables/12-deliverables-wp1**

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

-

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

As Above

Proposed corrective actions (if applicable)