VAMDC

Virtual Atomic and Molecular Data Centre

D7.3

–

Publishing Tools Report 2

Version 0.1

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:	D7.3
Deliverable title:	Publishing Tools Report 2
Due date of deliverable:	June 2011
Actual submission date:	August 2011
Authors:	U. Heiter, N. Piskunov, F. Kosmala, M.L. Dubernet and WP7 team
Work Package no.:	WP7-JRA2
Work Package title:	Publishing Tools
Work Package leader:	UU
Lead beneficiary:	UU
Dissemination level:	PU
Nature:	Report
No of pages (incl. cover):	

Abstract	The objective of D7.3 is to describe VAMDC Science/Technical Report for Cycle 2. This report corresponds to Activities in WP7: JRA2 "Publishing 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	**7 june 2011**	**Compilation of nodes reports for WP7 sent to leader**	**F. Kosmala**
V0.1	**July 2011**	**WP7 Report for P2**	**U. Heiter, N. Piskunov**
V0.1	**July/August 2011**	**Making of D7.3 document - Inclusion of above P2 Report in D7.3**	**F. Kosmala**
V0.1	**August 2011**	**Final Check of D7.3 Document**	**M.L. Dubernet**

Final Version (v0.1) released by		Circulated to	
Name	Date	Recipient	Date
M.L. Dubernet	24/08/2011	Mrs Asero	24/08/2011

WP7 ACTIVITIES DESCRIPTION

Work package number	7		**Start date or starting event:**				3			
Work package title	JRA2: Publishing Tools									
Activity Type	RTD									
Participant id	1	6	8	12	13					
Person-months per beneficiary: **(Total = EU + Node Contributions)**	12	12	12	5	24					

Table of Content

1. WP7 Objectives

JRA2 provides generic tools for A&M data in the VAMDC infrastructure. We envisage two publishing paths: (A) inclusion of the new data into existing databases, which are already integrated into the VAMDC, and (B) publishing data in a new database instance using an existing open source relational database software instrumented with the VAMDC interface.
The selection of specific path will be decided between the data producer and the VAMDC depending on the completeness of the new data, logical and structural similarity with existing databases, availability of the necessary resources at the producer site etc.

The process of publishing includes two steps: data quality assessment and the open offering of the new data to the VAMDC clients. In the first stage the new data will be technically included into the infrastructure but not made public. Instead, a group of experts working with the VAMDC will have access to the data in order to verify the quality assessment provided by the producer. After the quality assessment is completed the data will be offered in open access.

In order to implement this plan we need to formulate a number of procedures and develop or adopt a number of software tools. This work will not start from scratch: we will build on long experience accumulated by large databases handling many datasets from different data providers (e.g. VALD). In addition, several tasks in the WP7 are closely related to other work packages and thus require close and intensive collaboration.

WP7 Leader is UU(6).

2. WP7 Milestones and Deliverables

Milestones

M7.1	Technical meetings	WP7	UU	Months 5, 10, 16, 22, 28, 34, 40, 42	Minutes. Presentations on internal Website
M7.2	Evalation of softwares	WP7	UU	Months 10, 22, 34	

Deliverables

D7.1 Publishing Tools Plan (PM 3)
D7.2 Publishing Tools Report to be included in report to the EU – Year 1 (PM 10)
D7.3 Publishing Tools Report to be included in report to the EU – Year 2 (PM 22)
D7.4 Publishing Tools Report to be included in report to the EU – Year 3 (PM 34)
D7.5 Final Report of Publishing Tools to be included in final report to the commission (PM41)

Annual Publishing Tools Plan revisions included in Revised Annual VAMDC Project Plans – Year 1,2,3

3. WP7 Tasks Description

WP7 Leader (co)		
Task Number	Leader	Other Partners
1	M.-L. Dubernet/M. Doronin (CNRS)	All others
2	M.-L. Dubernet/M. Doronin (CNRS)	All others
3	N. Piskunov (UU)	All others
4	P. Loboda (RFNC-VNIIT)	All others
5	A. Fazlief (IAO)	All others

Description of work (possibly broken down into tasks)

This WP will develop software that will be deployed within the VAMDC infrastructure. Significant part of this software will be based on the standards developed in JRA1. The general software available to the VAMDC community will be accessible via the VAMDC technical website. Two alternative solutions are being developed. The first one offers a possibility to import new data to the existing informational resources, capable of importing data in standardized forms (tasks 2-4) developed within JRA1. The second option implies the

design of a generic information system accessible via the Internet (task 5) instrumented with the VAMDC integration tools developed in tasks 1-3. In this variant an automatic generation of semantic metadata for uploaded information resources is realized, taking into account the restrictions imposed by formal models of molecules and atoms. All software will be documented.

Task 1: Create/adapt tools to go from a DM/XML schema to a full database deployment with generation of automatic administrative interface. **(lead by CNRS(1) with (6))**

Task 2: Create/adapt tools to build registries from the content of databases **(lead by CNRS(1) with (6))**

Task 3: Create/adapt interfaces to easily update dictionaries **(lead by UU(6) with (1))**

Task 4: Develop software libraries using various languages allowing to easily generate output of already existing resources in standardized format (lead by RFNC-VNIIT(12) with (1), (6), (8))

Task 5: Create tools to upload, modify, retrieve, compare, visualize and publish information in molecular spectroscopy (lead by IAO(13))

4. WP7 Tasks Plans for Period 2

Year 2 for the WP7 is dedicated to the full implementation of the Publishing Tools plan finalized during the first year. We have established the two paths for publishing data in VAMDC: through existing VAMDC nodes and by deploying a new node. We have created and tested a prototype tool for importing data and for generating a new open-source database. We tested several open-source databases and selected MySQL as a recommended option although we will support other database software. While the Django-based tool offers plenty of flexibility for the format of the imported data the proto-type is not fully compliant with the VAMDC interface. The missing aspects are related to the work and standards established by other work packages and they are related to the dictionaries and registries. Thus the goal for WP7 in year 2 is to convert the prototype to a fully functional framework capable of importing and registering the new A&M data in VAMDC.

Task 1: Automatic generation of a new VAMDC node from a new data.

- Testing the prototype functionality with real-case data sets (VALD3 and Lund Atomic Data Centre, UU: Regandell, Marquart, Stempels, possibly CDMS and HITRAN).
- Converting the prototype into a production tool.
- Documenting the import procedures (UU: Heiter, Marquart, Stempels).

Task 2: Adding and testing the automatic registry update functionality
This work will be performed in close collaboration with WP6 and WP8.

Task 3: Dictionary manipulation and verification tools
Data import and registry integration require the publisher to describe the new data with two dictionaries: one related to the data content and the other to the query language functionality. We will provide tools for generating those dictionaries interactively with automatic compliance verification. (UU: Regandell, Stempels, possibly CNRS)

Task 4: Data presentation tools

The WP7 python interface based on the Django framework is one of a few interfaces developed within VAMDC JRAs. In this period we will test, compare and document these tools. This work will have the wide participation from other WPs and VAMDC nodes (UU, UNIVIE, INASAN, RFNC-VNIITF. KOLN, CNRS, QUB).

Task 5: Data quality control

The initial content of Task 5 aimed at creating tools to upload, modify, retrieve, compare, visualize and publish information in molecular spectroscopy is now complete (IAO). One of the important result is the automatic data consistency control based on quantum mechanical selection rules. In Year 2 we will investigate generalisation of this tool to in order to include it into the final WP7 product (IAO: Fazliev).

5. WP7 Tasks Reports for Period 2

Period: 01/07/2010 – 30/06/2011
WorkPackage: WP7/JRA2 Publishing Tools

WorkPackage Leader and co-Leader: N. Piskunov and U. Heiter (UU)

Participants in the WorkPackage: CNRS, RFNC-VNIITF, IAO, CMSUC, KOLN, INASAN, UNIVIE, QUB

Part 1

A summary of progress towards objectives and details for each tasks

Introduction

Individual tasks of this work package are closely related to other work packages through the selected data model, node software, registry content, query language and extraction tools. The main goal of Cycle 2 for the WP7 and other JRAs was to bring the software and data model from the proto-type level achieved in Cycle 1 to a fully functional and tested release level. A special effort was put into testing the Data model and Publishing Tools/Node Software with a variety of databases representing different data content and application fields.

The software development, tool and expertise sharing and reporting were organized as before, in a form of workshop series across all JRAs/SAs. During Cycle 2 we had three such workshops with the fourth planned for September 2011. INASAN organized the JRA-Coordination meeting of Cycle 2 in Nov 2010 in Moscow and participated in the JRA+SA Workshop 4 in Feb 2011 in Vienna, organized by UNIVIE. QUB organized the JRA+SA Workshop 5 in June 2011 in Belfast. For the contributions of the other nodes see task details below.

Task 1: Node Software development − generation of a new VAMDC node

(CNRS, UU, KOLN, CMSUC)

Data publishing in VAMDC is realized by setting up a VAMDC node, consisting of a database engine, a web server and a framework/interface providing the connection to the central VAMDC server (registry) and allowing standardized data queries. Two software packages which provide this functionality have been developed, implemented, and tested in WP7 during period 2.

The first one, called “NodeSoftware”, is based on the Python framework Django, and has been developed at UU. It allows to build a new VAMDC node from scratch, following the documentation provided on the VAMDC web site (http://www.vamdc.org/documents/nodesoftware). This NodeSoftware package has been deployed in period 2 for a number of existing databases (i.e. previously offering stand-alone web services): the VALD, CDMS, and CHIANTI databases by UU, KOLN, and CMSUC, respectively, as well as CDSD (IAO), UDFA (QUB), Spectr-W3 (RFNC-VNIITF), and the HITRAN, Ethylene, Methane and GHOSST databases. In addition, a completely new

VAMDC node was created in Lund from data distributed in various publications and unpublished data files.

LPMAA (CNRS) has been involved in developing a Java tool and libraries compliant with standards (VAMDC-XSAMS schema, Query language, Dictionaries) developed in WP6 in order to publish atomic and molecular data. This software package has been deployed for the BASECOL and KIDA databases, and had the advantage not to rely on any external platform. The tool understands the standard queries from VAMDC tools and returns data in VAMDC-XSAMS compliant format. It is an alternative to the Django platform developed at UU.

Task 2: Registry population/updating from database content

(UU in close collaboration with WP6 and WP8)

A semi-automatic registry update functionality is now included in the NodeSoftware package. NodeSoftware automatically reports its own version and the standards version it implements at a standard 'Capabilities' URL. It also reports lists of dictionary keywords called Returnables and Restrictables which tell what kind of data a node has and which queries will work, as well as example queries (for future automated testing). Node administrators (i.e. data publishers) can use the VAMDC Registry Web Administration tool to update the registry with this information (with only a few mouse clicks).

Task 3: Dictionary generation/updating

(UU in close collaboration with WP6)

The VAMDC dictionaries connecting the Node Software with XSAMS have been updated and we have developed a special tool for maintaining the dictionary database (see http://www.vamdc.org/documents/nodesoftware/concepts.html#the-vamdc-dictionary). The dictionary tool allows node administrators (i.e. data publishers) to generate dictionaries for their nodes consistent with the VAMDC-XSAMS data model, and provides a checking function for existing dictionaries. The dictionary tool contains a detailed description for many of the several hundred VAMDC-XSAMS keywords. Future updates will be made to maintain consistency with the XSAMS schema.

Task 4: Data import tool

(RFNC-VNIITF, UU, CNRS)

The data import process includes two steps: creating a set of external ASCII tables that match the database model and importing these tables using the LOAD DATA command of an SQL database. This approach has shown itself to be the most robust, efficient and easily adopted by the data producers. Thus the Import Tool is a stand-alone ASCII file manipulation routine written in Python capable of creating database-matching tables from a set of ASCII files. This *rewrite tool* requires a preparation of the "mapping file" describing the names and location of the input files, their layout and the layout of the output files. The software, documentation and examples can be found in the Node Software documentation (http://www.vamdc.org/documents/nodesoftware/importing.html). The rewrite tool has been extended according to the needs from its users and can now handle input data in a wide variety of formats.

In addition, the special-purpose online tool to export selected data from the Spectr-W^3 database to xml format conforming to the XSAMS schema, as well as to import XSAMS-

gauged xml data into Spectr-W^3 was improved and extended.

The work in this task during period 2 also lead to the detection of deficiencies in the schema related to data content in various databases, which became apparent during the implementation of the publishing tools. The results have been presented at the meetings in Cambridge in March 2011 and in Moscow in May 2011. Most of the suggestions were or will be implemented in the next release of the VAMDC-XSAMS and will be suggested to the IAEA XSAMS standard in October 2011.

An ingestion Tool has been developed at INAF in order to import results from Quantum Calculations (http://sourceforge.net/projects/qchitool/)

Task 5: Automatic data verification tool

(IAO)

An automatic tool for testing the formal consistency of the transition data sets from the perspective of quantum-mechanical selection rules was developed. The tool analyses an XSAMS data stream, collects a set of quantum numbers describing the initial and the final configurations and checks them against a list of selection rules. Such rules have been formulated for the HITRAN molecular transitions and for the VALD atomic data. The report generated by this tool can be used to identify and correct errors in data collections. The test implementation at the water-vapour spectroscopy database W@DIS at IAO (one of the VAMDC nodes) was extended from CO to 16 different molecules.

Significant results (Activities and Deliverables)

The activities are detailed in the box above. Significant results are

a) The efficiency of the data import tool has increased considerably. For example, the import of VALD data in period 1 required two hours of computer time, while the import of a new extended version of VALD at the end of period 2 could be done within 20 minutes.

b) Complete consistency of the publishing tools (and node software) with the VAMDC-XSAMS standard was achieved, in part through software adaptation, and to a great extent through VAMDC-XSAMS development in collaboration with WP6.

Deliverables to EU

D7.1 Publishing Tools Plan
See http://www.vamdc.eu/public-deliverables/19-deliverables-wp7

D7.2 Publishing Tools Report to be included in report to the EU – Year 1 – DONE – http://www.vamdc.eu/public-deliverables/19-deliverables-wp7

D7.3 Publishing Tools Report to be included in report to the EU – Year 2 – DONE – http://www.vamdc.eu/public-deliverables/19-deliverables-wp7

Annual Publishing Tools 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 Key internal deliverables were softwares for each of the 5 tasks, which are distributed on the VAMDC web site as part of the Node Software (at http://www.vamdc.eu/software) (Tasks 1-4), on the IAO web site (Task 5) [the general page is s http://wadis.saga.iao.ru/ and the direct link to the tool is http://wadis.saga.iao.ru/saga2/transition/comp2/) and on sourceforge for INAF tool as it is a prototype (http://sourceforge.net/projects/qchitool/)

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

n/a

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

n/a

Proposed corrective actions (if applicable)

n/a