



VAMDC

Virtual Atomic and Molecular Data Centre

D7.1

–

Publishing Tools Plan

Version 0.4

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	IVIC	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 supported by funding from the xxxxx Programme under the 7th Research Framework Programme of the European Union

Document

Deliverable number: D7.1
Deliverable title: Publishing Tools Plan
Due date of deliverable: September 2009
Actual submission date:
Authors: N. Piskunov, T. Marquart, P. Loboda, A. Fazliev, M.L. Dubernet
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 D7.1 is to describe VAMDC Publishing Tools Plan on PM3. This plan corresponds to Activities in WP7: JRA2 “Publishing Tools”. This plan is included in the VAMDC Project Plan.
----------	--

Versioning and Contribution history

Version	Date	Reason for modification	Modified by
V0.1	21/10/2009	WP7 Draft	M.-L. Dubernet
V0.2	30/12/2009	Milestones for T1-T5, List of responsibilities	N. Piskunov
V0.3	21/01/2010	Adjustment of dates, language corrections, INAF contribution	N. Piskunov T. Marquart M.-L. Dubernet P. Loboda
V0.4	22/01/2010	Detailed description for the Task 5	N. Piskunov A. Fazliev

Final Version (v0.4) released by		Circulated to	
Name	Date	Recipient	Date
N. Piskunov	22-01-2010	Mrs Asero	04-03-2010

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.

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

WP7 activities description.....	5
1. WP7 Objectives.....	5
2. WP7 Milestones and Deliverables	5
3. WP7 Tasks Description.....	6
4. WP7 Tasks Description for Period 1.....	7

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	Evaluation 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 Description for Period 1

The WP7 Work Plan includes final products of two types: formulation of a number of procedures and development/adoption of a number of software tools. Several tasks in the WP7 are closely related to other work packages and require close and intensive collaboration.

The procedures to be formulated for the Publishing Tools WP7 are:

- *Preparation rules for the original data.* We intend to offer reasonable flexibility for the format of data sets to be imported to a VAMDC database but the format must be restricted by small number of clear and easy-to-follow rules. These rules will ensure full description of the data (physical meaning, units, bibliography etc.) and the cross-referencing if the data comes in several parts. In the end we may support several formats but we start by developing a tool for importing data in form of ASCII tables. This document will also describe a unified DB interface and the methods of tuning it to the existing DBs.
- *Guidelines for the expert quality assessment.* This document will contain a summary of data quality assessment procedures adopted in major existing databases (e.g. VALD, NIST etc.) and a description of the data access protocol, available tools and data usage policy during the assessment period. This document will be revised in the future incorporating the VAMDC own experience, expert comments and suggestions and therefore it must be updated on a regular basis.
- *Selection of an open-source relational database to be used in Path B.* This document will contain a report explaining the motivation behind the selection of specific software.
- *Detailed deployment/verification/maintenance instructions for Path B.* Although the software and the VAMDC interface for Path B will be provided by the WP7, the actual deployment and data publishing must be thoroughly tested before adding a new resource to VAMDC. Both procedures must be described very clearly under the assumption that the data producer does not have special training in installing and running relational databases. A critical requirement for Path B is service availability and sustainability on a long run. These issues will also be addressed by this document.
- *Statistics and data bibliography.* We have identified the lack of proper referencing to the original data producers as the major concern for publishing in large databases. Solving this issue will be a major incentive for the data producer to publish their data through VAMDC. This document should specify the rules for storing and cross-referencing the bibliographic and A&M data as well as the logs of processed requests so any request can be traced back to the source of the extracted data.

Software tools to be developed in WP7:

- VAMDC interface for publishing new A&M data
- Data import tool
- Automatic data verification tool
- VAMDC interface for a selected open source database with the support of VAMDC data exchange protocols
- VAMDC interface for data quality experts (Quality assessment report)

Coordination and Cooperation:

Many of the items to be developed in WP7 are closely related or dependent on other work packages and thus require very close coordination and often collaboration to avoid effort duplication and interface incompatibility. While those questions will be in competence of WP2 we feel that it is our responsibility to identify them. The following specific topics of the WP7 involve interaction with other work packages:

- The format of internal data representation/transport in the VAMDC is critically important for the Data Preparation Rules.
- Guidelines for the quality assessment will be prepared by the WP7 but must be discussed and approved by the VAMDC VPB (WP1 and WP2). See e.g. the point on "quality assurance systems" in Task 1 of [WP2 plan for period 1](#).
- The selection of the open source database and its instrumentation is dependent on both the internal data representation and the registry query protocol adopted in VAMDC. This requires coordination with WP5 and WP6.
- Path B deployment document aimed at data providers must be coordinated with WP4 but also with WP3.
- The cross-referencing system and the request logging mechanism must be distributed between individual databases and the VAMDC server. This is as much policy question as it is a technical issue. Thus it must be a cooperation topic with WP2 and WP4.
- The data Import Tool interface to the VAMDC database(s) must comply with the VAMDC internal data representation. This calls for cooperation with the WP6.
- The software package for the Path B must be coordinated with WP4 and WP5.
- Automatic data verification tool overlaps with SA2:T5 of the WP5.
- VAMDC interface will be mostly developed within WP6 but parts of it will be included in WP7 while testing/verification will involve WP4 and WP5.

Coordinating activities will be carried out through systematic updating of the wiki pages, extensive email exchange, use of version control software for the toolkit development and regular working meetings. The following working meetings have been planned for Period 1:

1. Coordination with WP6, Köln, February 2010
2. XSAMS and registry toolkits, Paris, March 2010
3. Automatic registry updates and referencing system within VAMDC, Vienna or Florence, March 2010
4. Test deployment of the VAMDC interface prototype for VALD, Moscow (TBC), May 2010

Here is a description of specific steps within each task planned for Period 1:

Task 1: The current version of the XSAMS will be used as a basis for creating a toolkit for various VAMDC applications. The toolkit will be primarily aimed at data conversion to and from XSAM to ASCII tables and other formats ([April 29th, 2010, Doronin, Nenadovic, Dubernet](#)).

Task 2: A prototype toolkit will be created for collecting exhaustive information about the content of an A&M database, available extraction tools and procedures as well as the usage statistics ([June 4th, 2010, Doronin, Nenadovic, Dubernet](#)).

Task 3:

- Formulating the requirements for the "standard" VAMDC database software based on the analysis of the structure, content and selection tools of the existing databases (March 1st, 2010, Marquart, Heiter) Note: WP2 Task 1 includes a "census of all VAMDC resources" - the results of this task will be used in WP7/Task 3.

- Finalizing the WP7 plan, internal deliverables and time lines for each partner and writing Publishing Tools Report 1 (April 12th, 2010, Piskunov, Heiter).
- Preparing first version of specification for the format of the data to be imported into VAMDC (April 29th, 2010, Marquart, Piskunov).
- Preparing the prototype of the import tool (together with the RFNC-VNIIT, May 3rd, 2010, Marquart).
- Selection of the "short list" candidates for the VAMDC "standard" database (May 28th, 2010, Marquart, Stempels).
- Testing the import tool with VALD and Spectr W³ databases (June 11th, 2010, Marquart, Heiter, Stempels).

Task 4: Prototype development and deployment of the VAMDC XSAMS interface for the VALD and Spectr W³ databases (June 4th, 2010, Marquart, Laboda).

Task 5: A prototype automatic data verification tool developed for a "test" atom and CO-molecule will be created and tested for primary data sources. (May 28th, Fazliev).

- Formulating the requirements for the "standard" middleware based on the analysis of the atomic and molecular data and metadata structures and constructing first version of middleware (June 30, 2010, Akhlyostin)
- Preparing specification for the quantum numbers of CO-molecule, a "test" atom and the corresponding selection rules (March 1st, 2010, Perevalov, Ryabchikova)
- Preparing the prototype of the upload data system for the molecular data (CO-molecule transitions and line profiles data structures) from primary data sources (March 1, 2010, Lavrentiev, Privezentsev).
- Preparing the prototype of the upload data system for the atomic data (the atom transitions and line profiles data structures) from primary data sources (May 1, 2010, Lavrentiev, Privezentsev).
- Preparing the prototype of the tabular presentation system of molecular data (CO-molecule transitions and line profiles data structures) from primary data sources (April 1, 2010, Akhlyostin).
- Preparing the prototype of the tabular presentation system of atomic data (transitions and line profiles data structures) from primary data sources (April 1, 2010, Akhlyostin).
- Formulating the requirements for publication tools (analysis of the relations between the data owner and person reviewing the data). Preparing specification of these relations and prototype of the software implementing the publishing procedure. (May 28, 2010, Kozodoev)

Release of the software, documentation, test results, performance assessment and procedure documents for Tasks 1-5 will be available through links from the VAMDC web-site.