Wf4Ever
FP7-270192

Workflow 4Ever

PROJECT PERIODIC REPORT

Grant Agreement number: 270192

Project acronym: Wf4Ever
Project title: Advanced Workflow Preservation Technologies for Enhanced Science
Funding Scheme: STREP

Date of latest version of Annex I against which the assessment will be made: 29/11/2010

Periodic report: 1st X   2nd ☐   3rd ☐   4th ☐
Period covered: **from**   01/12/2010   **to** 30/11/2011

Name, title and organisation of the scientific representative of the project's coordinator: Dr. José Manuel Gómez Pérez, Intelligent Software Components (iSOCO) S.A.
Tel: +34913349797
Fax: +34913349799
E-mail: jmgomez@isoco.com
Project website address: http://www.wf4ever-project.org/

Project Information

This document is part of a research project funded by the IST Programme of the Commission of the European Communities as project number FP7-270192. The Beneficiaries in this project are:

Intelligent Software Components S.A. Edificio Testa Avda. del Partenón 16-18, 1º, 7ª Campo de las Naciones, 28042 Madrid Spain Contact person: Dr. Jose Manuel Gómez-Pérez E-mail address: jmgomez@isoco.com	**University of Manchester** Department of Computer Science, University of Manchester, Oxford Road Manchester, M13 9PL United Kingdom Contact person: Professor Carole Goble E-mail address: carole.goble@manchester.ac.uk
Universidad Politécnica de Madrid Departamento de Inteligencia Artificial Facultad de Informática, UPM 28660 Boadilla del Monte, Madrid Spain Contact person: Dr. Oscar Corcho E-mail address: ocorcho@fi.upm.es	**University of Oxford** Department of Zoology University of Oxford South Parks Road, Oxford OX1 3PS United Kingdom Contact person: Dr. Jun Zhao / Professor David De Roure E-mail address: {jun.zhao@zoo.ox.ac.uk, david.deroure@oerc.ox.ac.uk}
Poznań Supercomputing and Networking Center Network Services Department Poznań Supercomputing and Networking Center Z. Noskowskiego 12/14, 61-704 Poznan Poland Contact person: Dr. Raúl Palma de León E-mail address: rpalma@man.poznan.pl	**Instituto de Astrófísica de Andalucía** Dpto. Astronomía Extragaláctica Instituto Astrofísica Andalucía Glorieta de la Astronomía s/n 18008 Granada, Spain Contact person: Dr. Lourdes Verdes-Montenegro E-mail address: lourdes@iaa.es
Leiden University Medical Centre Department of Human Genetics Leiden University Medical Centre Albinusdreef 2, 2333 ZA Leiden The Netherlands Contact person: Dr. Marco Roos E-mail address: M.Roos1@uva.nl	

Table of contents

1. Publishable summary

1.1 Motivation and overall objectives

Wf4Ever[1] addresses some of the biggest challenges for the preservation of scientific workflows in data-intensive science, including: (a) the consideration of complex digital objects that comprise both their static and dynamic aspects, including workflow models, the provenance of their executions, and interconnections between workflows and related resources, (b) the provision of access, manipulation, sharing, reuse and evolution functions to these complex digital objects, (c) integral lifecycle management functions for workflows and their associated materials. To address these challenges, the Wf4Ever project investigates and develops *technological infrastructure* **for the preservation and efficient retrieval and reuse of scientific workflows in a range of disciplines.** In particular, Wf4Ever will enable:

1. The creation and management of complex Research Objects that take into account the dual nature (static and dynamic) of scientific workflows.
2. The archival, classification, and indexing of scientific workflows and their associated materials in scalable semantic repositories, providing advanced access and recommendation capabilities.
3. The creation of scientific communities to collaboratively share, reuse and evolve workflows and their parts, stimulating the development of new scientific knowledge.

Wf4Ever brings significant new functionality to scientists, enabling them to take a step forward in the preservation of scientific knowledge by introducing the **novel concept of workflow-related Research Objects**, which acknowledges the central role of workflows in e-science and their relevance for scientific preservation. We address the preservation requirements of scientific data by considering workflows as live entities, which as they evolve need to be kept consistent with respect to research materials, many of them beyond the control of the originating scientists. The **main tangible outcomes** of the project will be:

1. A **software architecture** for scientific workflow preservation and a **reference implementation** instantiating such architecture.
2. A new **Research Object model** for the description of scientific workflows and related materials.
3. New techniques and tools for **workflow decay analysis, abstraction and comparison**.
4. New techniques and tools for **Research Object evolution, personalised recommendations and collaboration between scientists**.
5. New techniques and tools for **integrity and authenticity management based on provenance models of workflow-related Research Objects**.
6. The application of our results and technology to two workflow-intensive scientific use cases in the areas of **Astrophysics** and **Genomics**.

Wf4Ever results are evaluated in real-world applications by tackling complex problems in the scientific domains of Astronomy and Genomics, including the preservation and management of scientific workflows from their very inception and the collaborative access and reuse by the scientific community to previously archived workflows. By carefully choosing these two applications, Wf4Ever aims to stimulate both the creation of methods and tools for the preservation of scientific data and the application of such advances in relevant scientific domains.

---

[1] http://www.wf4ever-project.org

1.2   Work Structure and Work Plan

Wf4Ever activities are divided in separate manageable work packages, while keeping a close eye on the overall system. The skills of the project partners are appropriate for carrying out the work proposed both from scientific and technological points of view. The next table and figure present the distribution of partners in work packages, as well as their leaders.

Table 1. Wf4Ever work packages and partner distribution

| WP	Name	Leader	Other partners involved
Technological Work Packages			
WP	Name	Leader	Other partners involved
1	Software Architecture and Technology	PSNC	OXF, All
2	Workflow Lifecycle Management	UNIMAN	iSOCO, UPM, PSNC, OXF
3	Workflow Evolution, Sharing and Collaboration	UPM	iSOCO, UNIMAN, PSNC, OXF
4	Workflow Integrity and Authenticity Maintenance	iSOCO	UNIMAN, UPM, OXF
Test Cases Work Packages			
WP	Name	Leader	Other partners involved
5	Astronomy	IAA	iSOCO, UPM, UNIMAN, PSNC, OXF
6	Bioinformatics	LUMC	iSOCO, UNIMAN, PSNC, OXF
Management Work Packages			
WP	Name	Leader	Other partners involved
7	Project Management	iSOCO	All
Dissemination, Transfer and Exploitation			
WP	Name	Leader	Other partners involved
8	Dissemination, Transfer and Exploitation	OXF	All
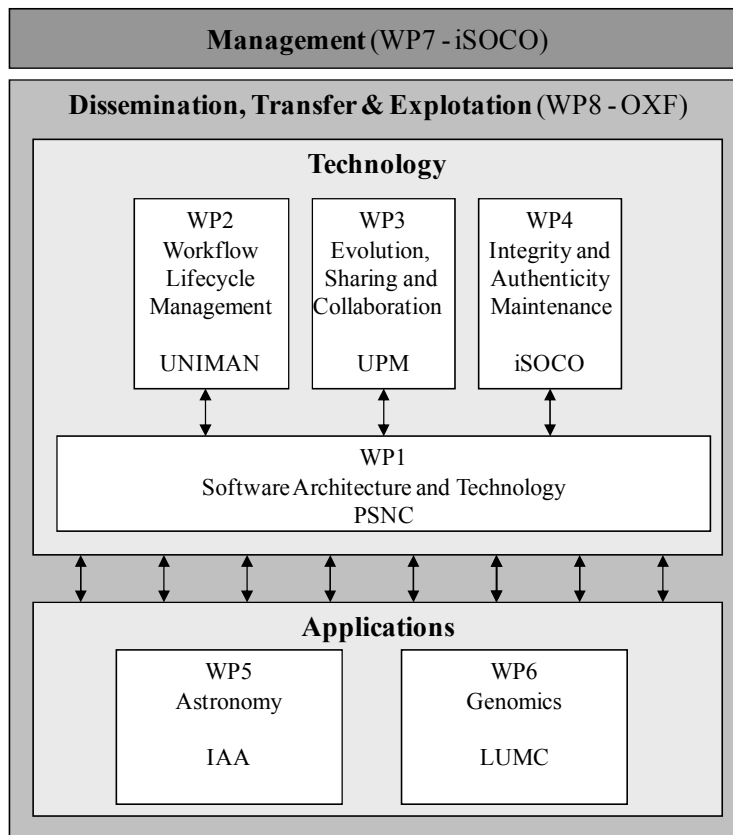
Figure 1 Overview of Wf4Ever Work Packages and leading organizations.

Wf4Ever is a **three year project** organized in **four phases** reflected in the milestones (please note that these phases overlap in months 33 and 34, since the final evaluations of individual components starts earlier than that of the use cases):

Table 2. Overall phases of the Wf4Ever project.

Phase 1 Months 1-6	Induction, know-how, and use case gathering. WP 1-4 will collaborate to assimilate the current architectural and technological features and challenges of all the areas involved; and identify appropriate approaches for their integration. WP 1-4 work with WP 5-6 to develop use case specifications that will form the drivers and the evaluation framework for the rest of the project
Phase 2 Months 7-22	A specification of the architecture incorporating all services from WP 2-4. A first compilation of all releases of the individual systems and of the integrated middleware, adequately evaluated throughout this period A first set of results on the application side
Phase 3 Months 23-34	A refined specification of the architecture incorporating all services from WP 2-4. A second compilation of all releases of the individual systems and of the integrated middleware, adequately evaluated throughout this period A second set of results on the application side
Phase 4 Months 33-36	Generation of final evaluation reports of individual systems, integrated platform and application use cases.

As a result of the kickoff meeting held in Madrid in December 2010, the consortium decided the best way forward to integrate resources and promote collaboration between project partners was to create horizontal task force organizations across work packages for each focused research and technical area. This led to four task forces:
- The Architecture task force.
- The Research Object and provenance modelling task force.
- The User task force.
- The Impact task force.

1.3 *Work performed during the first period and main achievements*

All milestones identified in the original workplan for the first year have been achieved, and all public deliverables are available in the project website. Therefore, there are no major deviations between the planning versus actual accomplishments. The main results obtained during this period are the following:
On the **application** side:
- Definition of the **initial requirements for workflow preservation and Wf4Ever tooling in the Astrophysics and Genomics domains**, including user profiles and demo scenarios that describe the expected results for the subsequent application development phases.
- **Characterization of the scientific workflows in Astrophysics and Genomics**. A compilation of existing standards, models and platforms, and similar initiatives has been performed.
- **Development of first workflows for Astrophysical Quantities Propagation and Genome Wide Association Studies of Metabolic Syndrome**. All workflows are available in

myExperiment[23]. Emerging Wf4Ever tooling was used in order to advance in the preservation aspects of these workflows.

- Provision of **continuous feedback to technical work packages** for the development of the notion of workflow preservation, Wf4Ever tooling, and the Semantic Model for Research Objects.

On the **technical** side, the following results have been also produced:

- The **Research Object model[4] v0.1**, which describes research objects as an aggregation of workflows and other resources. It comprises a suite of ontologies including ro (a skeleton for the aggregation of research object resources), wfdesc (a model describing initial descriptions of workflows), and wfprov (a minimal provenance model with links to other provenance models).
- A **Wf4Ever architecture** that identifies the most important services to be provisioned by the Wf4Ever middleware, and the interactions between those services. The architecture has been driven through the application use cases and will be validated by users in the next reporting periods.
- A **sandbox and integrated environment** where most of the basic software modules and libraries providing preliminary functionalities of the Wf4Ever tooling are already deployed.
- A **specification of the technical requirements** that need to be addressed by the technical work packages WP2-4 in accordance with the user requirements stemming from the application domains.
- Preliminary versions of the **Research Object evolution model** and **Recommender System**.
- A **characterization of Integrity & Authenticity for workflow preservation based on Provenance and Information Quality** dimensions and **early prototypes** based on a subset of such dimensions.

1.4   Expected final results and their potential impact and use

The project has had a high visibility through engagement with potential users of the project results, both in our application domains, Astrophysics and Genomics, where a good number of meetings with potential users have been held, and through contacts with other potential groups of early adopters. We have advanced towards the **creation of a user community around scientific workflows and their benefits for the Virtual Observatory community in Astrophysics**. An incipient group of interest has been formed in the international community, which has been federated into a specific **IVOA** (International Virtual Observatory Alliance) mailing list for workflows. On the other hand, **liaisons with user groups in Bioinformatics** and potential early adopters like the **Galaxy community** have also been initiated.

The project has been also presented in project concertation meetings, such as the one held in Luxembourg in May 2011 for "The Future of the Past – Shaping new visions for EU-research in digital preservation" workshop, organized by the EC.

Additionally, different outreach and communication activities with relevant stakeholders, including Libraries, Publishers, related FP7 projects, and standards have been made:

- **Librarians:** The objectives of the work being conducted in Wf4Ever were presented at the **Dagstuhl seminar**[5] in the workshop about The Future of Research Communication[6]. A

---

[2] http://www.myexperiment.org/packs/231.html
[3] http://www.myexperiment.org/packs/234.html
[4] http://www.wf4ever-project.org/wiki/display/docs/Research+Object+Vocabulary+Specification+v0.1
[5] http://drops.dagstuhl.de/opus/volltexte/2011/3315

collaboration with **DataVerse**[7], which is focused on supporting publishing, citing and discovering research data, has also been initiated. Finally, members of the Wf4Ever consortium participate in the **Concept Web Alliance**[8] on the preservation of scientific data.

- **Pusblishers,** including **Elsevier** and **BGI**.
- **Collaboration with other related FP7 projects** like the **SHIWA** has also been initiated whose formalization is expected in the form of a memorandum of understanding. Collaboration has also been established with the **SCAPE**[9] and **BioVel**[10] projects.
- **Standards:** The consortium has been working actively in the coordination with standardization efforts, such as the **W3C Provenance Working Group**. The consortium has also been in contact with **the Annotation Ontology**[11] (AO) team and participates in the **W3C Open Annotation Community**[12] Group. Consortium members also participate in the **Semantic Web Health Care and Life Science Interest Group.**

Papers in international conferences and workshops from the work performed during this first year will start appearing mainly during year 2.

[6] http://dx.doi.org/10.4230/DagRep.1.8.29
[7] http://thedata.org/
[8] http://www.nbic.nl/about-nbic/affiliated-organisations/cwa/introduction
[9] http://www.scape-project.eu/
[10] http://www.biovel.eu/
[11] http://code.google.com/p/annotation-ontology/wiki/Homepage
[12] http://www.w3.org/community/openannotation