Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science

STREP FP7-ICT-2007-6 270192

Objective ICT-2009.4.1 b) – "Advanced preservation scenarios"

D3.1: Workflow Evolution, Sharing and Collaboration Initial Requirements

---

Deliverable Co-ordinator: Rafael González-Cabero (UPM)

Deliverable Co-ordinating Institution: Universidad Politécnica de Madrid (UPM)

Other Authors: Raul Palma (PSNC), Carlos Ruiz (iSOCO). Khalid Belhajjame (UNIMAN)

This is an abstract of generic template for the Wf4Ever project, which shall contain all the necessary styles to help you produce a reasonably and consistently looking deliverable.

Document Identifier:	Wf4Ever/2011/D3.1/0.1	Date due:	M6
Class Deliverable:	Wf4Ever 270192	Submission date:	31/5/2011
Project start date:	December 1, 2010	Version:	v1.5
Project duration:	3 years	State:	Final
		Distribution:	Public

1 Wf4Ever Consortium

This document is a part of the Wf4Ever research project funded by the IST Programme of the Commission of the European Communities by the grant number FP7-ICT-2007-6 270192. The following partners are involved in the project:

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	

2 Change Log

Version	Date	Amended by	Changes
1.1	04-04-2011	Rafael González-Cabero	Initial version
1.2	20-5-2011	Carlos Ruiz	Addition of Collaboration Spheres requirements
1.3	21-5-2011	Raul Palma	Evolution and collaboration model requirements update.
1.4	27-5-2011	Rafael González-Cabero	Requirements table update. QA related fixes. Minor errata corrections
1.5	30-5-2011	Rafael González-Cabero	Minor corrections

3 Executive Summary

This document presents the initial set of requirements for the Research Object collaboration and evolution model; the Research Objects recommender system; and the sharing infrastructure that uses Collaboration Spheres visualization. The aim of this model and software is the provision of adequate means to maximise share and reuse of the preserved Research Objects, while supporting their evolution and versioning in order to support facilitating collaboration among scientists. These features will allow us to overcome the limited collaborative support for sharing and reusing workflows that exists in most of present workflow management systems and repositories.

This document contains the initial list of requirements and provides a starting point for design and implementation work. Nevertheless, as we propose a user-centric agile software development approach, this document should be considered a living document; requirements will change over time throughout the life of the project.

4 Table of contents

5 List of Figures

6 List of Tables

1. Introduction and Motivation

Scientific dissemination is still mostly based on the traditional notion of peer reviewed paper publication. This model encourages authors to write as many papers as possible; causing dissemination overhead for themselves and for the community of reviewers; but more importantly, to their colleagues and possible present and future interested target audience. Besides, all the generated information is made for human consumption, and as such textual which usually means unstructured and ambiguous content. This situation is even more flagrant since the widespread adoption of *in silico* experiments in many fields of science. Experimental or observational data, the associated models, and procedures specified as workflows are completely digital; they have no physical counterpart. Machine-interpretable experiments results, standard operating procedures and algorithms for analysis or simulation could thus be effectively be shared following the Linked Data principles [1] . Moreover, instead of static results written black over white on a piece of paper, researchers could be provided with machine-interpretable but human-readable, reproducible, reusable and traceable scientific content.

The evolution of paper publications is still early stage of development; they are just refurbished PDF versions of their paper counterparts (with just the addition of some lexical tags in the best cases). Resuming [22] computers cannot deal with information as produced in the classic articles electronic versions as they have the undesired properties such as ambiguity, lack of structure, and inaccessible and spread data. Data used in such publication is either not published (e.g. negative studies), not freely available, or not easy to find [22] .

In order to provide an alternative to traditional publications the Research Object concept has been coined. Research Objects are semantically rich aggregations of resources [2] that bring together data, methods and people in scientific investigations [3] . As described in [3] their goal is to create a class of artefacts that can encapsulate our digital knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge. We will focus in workflow-centred Research Objects, a subtype of Research Objects which methods are implemented as scientific workflows. These multidimensional digital objects comprise meaningful workflow models, the provenance of their executions, related services and datasets, documentation, etc.

The main objective of WP3 is to provide adequate means to maximise the share and reuse of the preserved workflow-centric Research Objects, while supporting their evolution and versioning, enabling the collaboration among scientists. In order to achieve this objective we propose a social approach that takes advantage of social features stemming from social networks and complement them with formal models, while considering Research Object evolution at the core of the proposed model. Our research will produce three main outcomes:

- ***Research Objects collaboration and evolution model***. Scientific knowledge dissemination nowadays is based on the traditional notion of paper publication, which quality and relevance is based on a peer review process. As described in [26] this approach encourages authors to write many papers as possible to get more “tokens of credit”. Whenever a progress is made on a certain subject, a new paper is written, reviewed, and published [26] . Henceforth, there is a lack of evolution traceability, nor support or encourage for reusing or collaborating in the development of publications. In the context of this work package we aim at proposing a Research Objects collaboration and evolution model that solves such problems. It will provide a precise description of the evolution tracing precisely the progress of a Research Object; and it will enable also the collaboration among researchers in the creation of Research Objects, providing means for describing mixed stewardship situations.

- ***Research Objects recommendation system***. Despite the many advantages of handling Research Objects over traditional static publications, the possibility of overwhelming researchers with too much irrelevant information, causing an information flood, is still present. The chasm between data production and data handling has become so wide, that many data go unnoticed or at least runs the risk of relative obscurity [27] . There is a high risk that valuable information could end unnoticed whilst researchers are overwhelmed with peripheral information (what is known as the buried in papers problem). Simple search mechanisms are not suitable, as they are passive and they presume that the user has at least the notion of the existence of the searched scientific item, which in cases such as a newcomer to a research field is not always the case. Instead, we propose the creation of a recommender system. Contrary to the search activity, recommender systems bring useful hints to the researcher in a proactive manner providing, without prior request by the user, practical suggestions of scientific data and results even though the user might not be aware of their existence beforehand. We will propose novel multi-faceted recommendation techniques; which will mix community-based and advanced demographical/content-based recommendations along with collaborative filtering techniques, which permit the discovery through collections of Research Objects
- ***Research Object sharing by means of Collaboration Spheres***. Since research in all fields is increasingly collaborative and a Research Object implies the aggregation of a number of workflows and resources, some of these resources will come from previous research work (even aggregated to other available Research Objects). Connecting Research Objects through the resources they share will create research networks of scientists and research groups in term of their interest (see Figure 1). A particular research network creates a collaboration sphere around those shared resources when scientists share their resources (i.e. a data set) with others so that they all can benefit from the publication of data (and even the analysis, for example).

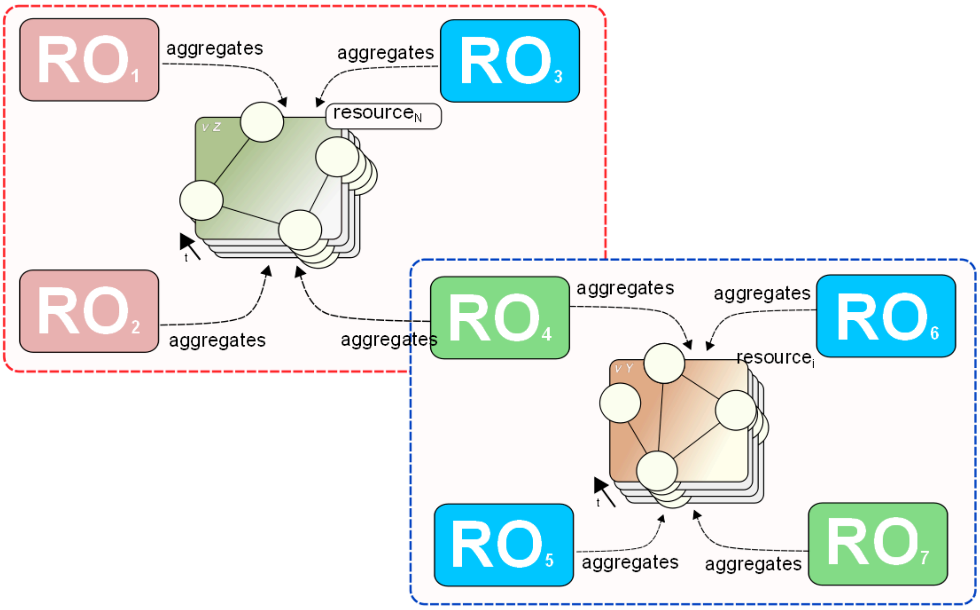

Figure 1 – Collaboration spheres

Besides, understanding the relationships between research objects, resources, scientists, and research groups is difficult since means a huge number of elements and relationships of different types (including those for supporting aspects of authenticity and integrity management). The situation requires effective visualization tools to build the graphical networks of scientists, resources, and workflows and explicitly representing the relations between scientists and their organizations.

Moreover, in some situations, scientists are not aware of the available Collaboration Spheres related to the work-at-hand. On those cases, they need filtering mechanisms to support advanced search based on semantic similarity and social analysis of the available Collaboration Spheres.

2. State of the Art

6.1 Evolution and Collaboration

The conceptualization of compound digital objects of scientific nature has already been investigated in the past. Previous works in this area have proposed specifications for characterizing and managing such compound objects, and in some cases they have produced abstract models for their representation. However, with respect to the evolution and collaboration aspects of those objects, less work has been done so far.

In Driver-II project, the concept Enhanced Publications has been defined as compound digital objects that combine ePrints with one or more metadata records, one or more research data objects, or any combination of these. The term ePrint is used to refer to an electronic version of an academic research paper, whereas the term research data may refer to different types of digital objects of scientific nature, such as: data collections containing, e.g., the results of experiments or measurements, data visualizations, multimedia files, text documents, software, comments and annotations, specifications of instruments, etc. In [11] , authors identify a set of requirements for storing and managing enhanced publications, which are the basis for the proposed data model. Among those requirements, we briefly introduce those related to evolution and collaboration. Regarding evolution aspects:

- It must be possible to keep track of the different versions of both the enhanced publication as a whole, and of its constituent parts. In particular, agents should be able to refer to specific versions of the compound object and its individual components. They propose that versions can be identified recording the date of the last modification, some version id, or a textual description of the version.
- In order to ensure the long term preservation of enhanced publications, it would be important to capture versioning information that will allow to decide to preserve one specific version of the enhanced publication, instead of having to wait until the entire enhanced publication is complete
- They argue that versioning information is an important relation between the web resources that are part of an enhanced publication, which must be captured.

Regarding the collaboration aspects, authors argue that it must be possible to capture the authorship of the enhanced publication and that of its constituent parts. This requirement is derived from the fact that e-Science projects are increasingly collaborative and interdisciplinary processes, and as a result it should be possible to trace individual contributions. This information in turn will also help to establish the trustworthiness of the resource.

The data model they developed considers those aspects as well as the other requirements identified. This model considers five main entities (ePrints, data objects, metadata, compound datasets and enhanced publications) and their key properties. It also considers the possibility to keep track of the different versions of both the enhanced publication as a whole, and of its constituent parts, to capture the provenance of the enhanced publication and of the various resources that it combines, e.g., authorship information, and to describe relations, such as versioning, between resources.

They also propose a number of vocabularies that can be used to make enhanced publications semantically interoperable using standardised and controlled vocabularies as much as possible. For instance, the DCMI Type Vocabulary provides a number of terms that may be used to describe the semantic type. Containments relations, relations between different versions, digital manifestations, bibliographic references and usage rights can be stated explicitly by making use the Dublin Core Metadata Initiative. To describe lineage relations, the ABC model may be used.

The OAIS information model [23] describes the functional components of an OAIS-type archive and provides a high-level description of the information objects managed by the archive. It built is around the concept of information package: a conceptualization of the structure of information as it moves into, through, and out of the archival system. It distinguishes three types of objects, submission, archival and dissemination information packages; being the second (AIP) the most important as it is focussed on the long-term preservation of the object. Among the metadata of an AIP, provenance information concerns about evolution aspects of the object. It documents the history of the Content Information, including its creation, any alterations to its content or format over time, its chain of custody, any actions (such as media refreshment or migration) taken to preserve the Content Information, and the outcome of these actions.

The LiquidPub[1] project proposes a new paradigm in the way scientific knowledge is created, disseminated, evaluated and maintained. This paradigm is enabled by the notion of Liquid Publications, which are evolutionary, collaborative, and composeable scientific contributions. In order to do so, they define the structure and evolution of Scientific Knowledge Objects (SKOs) [8] . They identify four layers to capture the particular aspects of the scientific artefacts they encode: file layer – concerns general content information, e.g., text, pictures and tables; semantic layer - includes all metadata and relational information related the objects from the previous layer; Serialization layer - enables and tracks aggregation and reuse of the information from the previous two layers; presentation layer – enables detaching visual format and display consideration from the actual content and meta-information. Moreover, they propose a state-based evolution model based on characterizing artefacts as belonging to the Gas, Liquid and Solid states, and assigning them properties based on those states. They also discuss how some aspects of version control can be implemented using semantic annotations, e.g., *isVersion*, *isSplit*, *isMerge.* Regarding the collaboration aspects, they discuss how the SKO model supports information about licensing and credit attribution tracking author the authorship of objects. They propose an XML notation to represent all those concepts. Finally, in a related thesis [30] , the author analyzes the problems and challenges of defining and modelling communities, and proposes among others a conceptual model for the definition, discovery, maintenance and use of such communities, and uses a relational schema to implement it.

Some other relevant models that address some aspects of evolution and collaboration are:

- *myExperiment data model.* It focuses on the modelling of content management (e.g., contributions, uploads), social networking (e.g., users, groups, friendship and membership) and object annotation (e.g.,

[1] http://liquidpub.org/

citations, comments, ratings, reviews and tagging) aspects especially around scientific workflows and experiment plans.

- *Admiral project*. It aims at facilitating the capture of research data and its subsequent publication via an institutional repository, has proposed data packages as transferrable entities (or "object") that contain both data and metadata describing that data. The structure of these packages, implemented in RDF/XML format, considers simple information about the evolution of the object, e.g., version, date of creation and date of last modification.

From the discussion above, we can see that there are already some works dealing with versioning of scientific (complex) objects, mainly focussing on the identification of versions and their relationship (esp. enhanced publications and liquid publications). In particular they propose simple models and the usage of basic annotations for handling those aspects. However, the management of changes from one version to another is one aspect that has not been explored yet. This topic has been explored for other relevant objects and models, such as ontologies. For example, Stojanovic's [44] and Klein's [44] propose a similar classification of changes in ontologies (atomic or elementary, composite and complex), where atomic changes refer to operations at the entity level. Some other works in the literature resemble (e.g., [18] and [25]) or extend (e.g., [19] and [28]) the previous models or provide partial solutions (e.g., [28] , [16] and [12]). More recently, in [31] , the authors propose a layered model for ontology changes that consists of a generic change ontology independent of the underlying ontology language. It also considers a more granular classification of changes (atomic, entity and composite) and the modelling of information for keeping track of the different versions.

Similarly, with respect to the collaboration aspects during the creation and maintenance of scientific objects, previous works have already considered the modelling of social information, such as communities (e.g., [30]), or social networking aspects (e.g., myExperiment). Still missing is a deeper consideration of the collaboration process and related aspects, such as roles and associated actions. This has also been addressed in the case of related objects and models, such as ontologies. Some early conceptual efforts were presented in [46] and [10] . Also, the work from [39] —and derivative works— provided recently a proposal, although without any technological support. Similarly in [30] , the authors proposed more recently conceptual model with technological support the process followed by organizations for the coordination of the change proposals. They propose to formalize this process by means of a collaborative editorial workflow model, which was implemented in a workflow ontology.

6.2 Recommender Systems

Recommender systems were originally defined as those systems that people provide recommendations as inputs and the system then aggregates and directs to appropriate recipients [36] . It is so because the first recommender systems, firstly Tapestry [13] and then others, were rooted in the use of the collaborative filtering technique. Recommender systems definition was later generalized, becoming a recommender system any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options [7] . In Table 1 the main basic recommendation techniques are represented, together with a summary of their main advantages and drawbacks.

Technique	Advantages	Drawbacks
Collaborative filtering	Can identify cross-genre niches Domain knowledge is not needed Adaptative: quality improves over time Implicit feedback sufficient	New users handling New item handling Quality dependent on large historical data set Gray sheep problem Stability vs plasticity
Content-based	Domain knowledge is not needed Adaptative: quality improves over time Implicit feedback sufficient	New user handling Quality dependent on large historical data set
Demographic	Can identify gross-genre niches Domain knowledge is not needed Adaptative: quality improves over time	New user handling Quality dependent on large historical data set Gray sheep problem
Utility-based	No cold start problem Sensitive to changes of preference Can include non-product features	User must explicitly define the utility function Static
Knowledge-based	No cold start problem Sensitive to changes of preference Can include non-product features Can map users needs to products	Static Knowledge engineering required

Table 1 Basic recommendation techniques advantages/drawbacks (adaptation of [7])

- ***Collaborative filtering***. Collaborative filtering techniques predict user's affinity for items on the basis of the ratings that other users have made to these items in the past. Therefore, the steps taken to make recommendation in such systems consist in finding people with similar tastes to the user by means its past ratings; and by means of their ratings extrapolate the user future ratings. User information in a collaborative system consists of a vector of items and their associated ratings; finding similar users translates into finding similar vectors. The main advantage of collaborative techniques is that they are completely domain independent; they treat objects of any complexity as black boxes that are analyzed in a crowd-sourced way. This approach enables their capability to identify cross-genre niches, discovering relationships between users and items that might not be apparent. Moreover, their functioning improves over time, as the system posses more precise knowledge about user set tastes and interests. Their most important disadvantages are so called cold start problems, namely the new user handling and the new item handling. Collaborative filtering recommender systems rely in historical information; handling addition of new elements (either a new user or recommendable item) that are neither reflected nor referenced in this background knowledge causes problems.

- ***Content-based***. Content-based recommender systems (e.g. [4] , [24] , [38]) make use of information retrieval and filtering techniques. A content-based recommender tries to infer users future items of interest on the basis of the features of the objects that the users rated in the past. These object features are items of interest such as keywords that define the object, a summary of its content, etc.. Content-based techniques have similar advantages to collaborative filtering approaches (without the ability of detecting cross-genre niches), and they do not exhibit the new item problem. Nonetheless, they still rely in a large historical data set.
- ***Demographic.*** Demographic recommenders (e.g. [37] , [21] , [33] etc.) systems use machine-learning techniques to cluster users on the basis of their personal attributes. Recommendations are based on these demographic classes, associating items with the class of users that employ them. The advantage of demographic techniques is that they do not rely on historical information of user past preferences, therefore they lessen the cold start problem. Their main drawbacks are most of those of content-based systems plus the necessity of handling personal (and possibly sensitive) users information.
- ***Utility-based***. As described in [7] utility-based (and knowledge-based recommenders as we describe later on) do not attempt to build long-term generalizations about users, but rather base their advice on an evaluation of the match between users needs and the set of items available (e.g. [15]); making suggestions based on a computation of the utility of each object for the user. The benefit of utility-based recommender systems is that they can handle complex functions that take into account different attributes not necessarily contained in users/items descriptions. Their main disadvantage is the difficulty of defining a proper utility function; and once defined it is hard to change, making these systems somehow static.
- ***Knowledge-based.*** Knowledge-based techniques (e.g. [48] , [42] , etc.) attempt to suggest objects based on inferences about a user's needs and preferences [7] . These systems handle knowledge about how a given item may solve a particular user need; and they make recommendation reasoning with this knowledge. Their main advantage is that they do not rely on historical information, and therefore there is no new-user/new-item problem. These systems are also very sensitive to changes of user preference, as they are articulated in an explicit manner; and as utility-based systems, can include additional features external to items in order to improve recommendations. Nevertheless, their main drawback is similar to the drawback of utility-based systems, they are very difficult to set up and tend to be too static.

In order to lessen the main drawbacks of each of these techniques, and as an attempt of obtaining several of the unique benefits that each technique brings, several hybrid approaches have been proposed (e.g. [49] ,[43] ,[41] ,[33] , etc.). These hybrid approaches make use of some (or all) of the above-described techniques, combining them, as was presented in [7] , following one of these methods:

- ***Weighted***. The scores given to an item by different recommendation techniques are combined together to produce a single recommendation. Each technique receives a weight (either static or dynamic if the hybrid approach takes into account the applicability of each recommendation technique to the concrete situation).
- ***Switching***. The system switches between recommendation techniques using the more suitable for the current situation.
- ***Mixed***. Recommendations calculated from several different recommenders are presented at the same time to the user as a multidimensional result.
- ***Feature combination and augmentation***. Features from different recommendation data sources are tailored and combined together for being used by a single recommendation algorithm. In the case of the feature augmentation approach, the output from one recommendation technique is used as an input feature to another.
- ***Cascade***. One recommender refines the recommendations given by another. This combination technique is especially useful for mixing utility-based and knowledge-based filters with other recommendation techniques.

6.3 Collaboration Spheres

The term policy has been used in the literature in a very broad sense, referring from security policies or trust management to business rules [34] . In general, two main areas within the topic can be identified:

- Security policies that rely on strong security mechanisms, based mainly on user identity and authentication e.g., trusted certification authorities.
- Trust management systems that rely on procedures for establishing and maintaining trust relationship among users in large open systems where anyone can contribute and access in somehow.

The main purpose of policies is to be able to dynamically control and automate the behaviour of complex environments without requiring code changing or the global cooperation of all system components. In [6] , the main benefits of this approach were identified: reusability, efficiency, extensibility, context-sensitivity, verifiability, support for both simple and sophisticated components, protection from poorly designed components, and reasoning about component behaviour. In a broad sense, this approach encompasses [6] [5] :

- Access control and privacy policies protect any system open to the Internet and assist users while they browse and interact with different resources and services. In general, Security Policies pose constraints on the behaviour of a system (e.g., used to control permissions of users and groups while accessing resources).
- Network administration policies are often applied to automate and govern network administration tasks, such as configuration, security, recovery, or quality of service.

- Trust Management policy languages are used to collect user properties in open environments, where the set of potential users spans over the entire Web.
- Action Languages are used in reactive policy specification to execute actions like event logging, notifications, etc.
- Business Rules are statements about how a business is done and are used to formalize and automate business decisions as well as for efficiency reasons.

In order to regulate access control permissions, monitoring, and other actions to be taken, policy-based systems generally rely on strong security mechanisms such as signed certificates and trusted certification authorities (CAs).

Policies have been specified in many different ways, and multiple approaches have been proposed. In the context of this deliverable, we concentrate on three well-known web languages for policy representation and reasoning, KAoS [50] [51] , Rei [17] , and Ponder [9] , and provide a brief summary on their main features below (while a more complete comparison is described in [47]):

- KAoS is a set of independent-platform policy service that allows the creation, management, and conflict resolution for policies, providing the capability for groups of people, resources and any other entity to be structured into domains within organizations. One of its interesting features is that policies are represented as ontologies in OWL, distinguishing between authorization and obligation for an action to be performed. Other policy actions, such as role-based authorization, are built from the basic domain primitives plus the policy types. Additionally, the KAoS framework allows the use of additional domain ontologies to express related concepts and actions, and it supports dynamic runtime policy change and update due to a logical inference engine which resolves policy selection and conflicts at runtime. The framework also defines ontology-based mechanisms to load new platforms and applications. A sophisticated graphical tool called KAoS Policy Administration Tool (KPAT) that facilitates security designers to focus on high-level policies specification, visualization and monitoring is also available.
- Ponder is a declarative and object-oriented policy language for distributed systems developed at Imperial College, which covers concepts like domains and roles in a organization, and relationships to groups the object to which policies apply. It also supports obligation policies in form of condition-action rules, registration of users and audio event for security violations. A complete toolkit has been developed to support the users of the language (a grammar compiler, a policy editor, and a management toolkit).
- Rei is a policy language based in OWL-Lite developed by L. Kagal, concerned with support for pervasive computing applications. In Rei, policies can be specified as constraints over allowable and obligated actions on resources. It also includes meta-policy specifications for conflict resolution and policy analysis specifications like what-if analysis and use-case management. It offers a reasoning engine to provide answers about the current permissions and obligations of an entity in order to guide its behaviour.

3. Methodology

In this section, we summarize the methodology we have followed to extract the set of initial requirements from bioinformatics and astronomy domains for the Research Objects collaboration and evolution model, the recommender system, and the Collaboration Spheres based Research Objects sharing system. We have carried out the following steps:

- ***Definition of User scenarios***. We begin by defining user scenarios (which we refer as golden exemplars) from our target users' domains of activity. Users should articulate their goals and requirements in a technological agnostic manner, focusing in providing a neutral description of their day-to-day work. Users should focus on what they would like to achieve. It is also important that users first capture what is important to them, rather than produce a comprehensive list of everything that they might one day find useful.
- ***Isolate user requirements***. From the user scenarios and golden exemplars we distil a set of user requirements. This involves a close examination of the user supplied materials, and extracting any information that can be interpreted as a goal or requirement of the user in their day-to-day work, and any a benefit they may realize by virtue of having any such requirement satisfied. We follow a common agile development practice for describing user requirements, user stories. Firstly we propose a set of representative user roles that represent significant user types for each use case; them we extract users desires expressing them in the form: "As a [user role] I want [articulation of requirement] so that [description of ensuing benefit to the user]". This form helps to focus attention on a user's needs rather than the technical means whereby they might be satisfied.
- ***User review***. The distilled requirements are shown to the users to ensure that the original scenario has been properly captured, and therefore they properly reflect the scenario described, or to clarify any misunderstandings there may be in its interpretation.
- ***Cluster requirements***. Requirements extraction step described above focuses on extracting maximum requirement information from the user-supplied materials, and can result in many similar or overlapping requirements. In this step these extracted requirements are examined and grouped into similar or overlapping clusters. For each of these groups, we re-articulate the goals and benefits expressed as one or more requirements, ideally non-overlapping. This process should result in a reduced number of requirements that can be treated relatively independently of each other.
- ***Project technical requirements***. In this step, user requirements are assessed in the context of a technical deployment environment, and corresponding technical requirements are proposed so that user desires can be satisfied. At the same time, assumptions about the nature of the technical environment should be articulated and provide a basis for justifying technical design decisions.
- ***Classify technical requirements.*** In this final step we organize the technical requirements into different categories. We have classified them in the following groups of technical requirements:
    - ***Research Object Dimensions Requirements.*** These technical requirements are those that are closely related with the different dimension and properties of Research Objects. These

properties and dimension, which are based on some of the R's described in [3] , and later extended on [2] , focus primarily on reuse; describing the ways in which information within a Research Object is, or might be, reused (and how that reuse might occur). Nonetheless, other concerns of great importance such as provenance, evolution, consistency, etc. are also covered. We remit the reader to deliverable D2.1 Workflow Lifecycle Management Initial Requirements, where the whole description of the dimensions is included.

- ***Projected User Requirements.*** These requirements are projected almost unaltered from the users requirements identified in user stories. The difference with Research Object dimensions requirements is that they have no relation with the properties identified for Research Objects, but nonetheless, they must be taken into account (and perhaps might result in future properties of or functionalities around Research Objects).
- ***Activity Specific Requirements.*** Though not explicitly extracted from user stories, these requirements further restrict the characteristics of the target system in order to provide the expected functionality to the user. They cover specifics related with research concerns or issues that are well identified in the sate of the art of the activities covered in this document. These activities are the definition of the evolution and collaboration Research Object model, and the provisioning of mechanisms to allow the automatic recommendation and sharing of Research Objects.
- ***Low-level requirements.*** Low-level requirements depict specific details about the final implementation of the system. As such, they should be transparent to the user, and they should neither restrict neither interfere with any of the user-related requirements. The main sources of low-level requirements are standard and technological compliance issues, and the description of work of Wf4Ever.

Finally, each of these sets of requirements has been divided in evolution and collaboration model requirements, recommendation requirements, and sharing requirements.

4. Use cases and user requirements summary

In this section we provide a brief but illustrative description of the use cases. For a full in detail description we remit the reader to the deliverables D5.1 Astronomy Workflow Preservation Requirements, for a complete description of the workflow preservation in astronomy use case (astronomy use case for short); and D6.1 Genomics Workflow Preservation Requirements for a complete description of the workflow preservation for genome wide analyses for genomics and bioBanking communities (Bioinformatics use case for short). In our summary we include a brief description of the use cases (focusing in the golden exemplars that users have identified); the different user's roles that we have distinguished from these golden exemplars, and the user stories that better illustrate users' desires.

6.4 Astronomy use case summary

Astronomy is among the first scientific disciplines to embrace and benefit from early development of web-based technologies enabling cross-linking of resources across archives. Our partners from the Instituto Astrofísica de Andalucía have identified three golden exemplars that are representative of the experiments that are performed within the astronomy field. Briefly they are:

- ***Propagation of quantities***. The first golden exemplar showcases the need to update values that are dependent on other volatile values. Specifically, the user is interested in maintaining the freshness of data values that measure the magnitude, distance and intrinsic luminosities of a set of objects. The process, by which the freshness of such values is maintained, is implemented using a workflow. Such workflow is enacted every time values of variables, on which the magnitude, distance or intrinsic luminosities depend, are updated.
- ***Extraction of galaxy samples***. The second use case depicts how users how users retrieve a set of 2D images from existing catalogues with the objective of identifying a list of potential objects, e.g., companion galaxies and their hosts, that meet given special distribution criteria in the sky.
- ***Modelling of 3D data of galaxies*****.** The last golden exemplar illustrates the need for processing and transferring large volume of data, generated and processed using workflows, in the form of 3D binary cubes. Two dimensions of these cubes correspond with the spatial dimensions, whereas the third corresponds with the velocity of the gas emitting the light captured.

The above-described golden exemplars were used to elicit many requirements, which will be described later on in this document. It is however worth mentioning at this stage that with regard to the activities that are the focus of this document, the main requirements are: (i) the need of providing mechanisms to better allow the sharing and discovering workflows and their related scientific data and procedures (such as datasets, experiments logs, scripts, services descriptions, and a long etc.); and (ii) the need to provide methods for versioning and tracing the evolution of such workflows, procedures and scientific data (even in the case where multiple parties are involved in a collaborative fashion)

6.5 Bioinformatics use case

One of the main issues in biomedical research lies in the study of large datasets, and combinations of thereof, with the objective to understand the mechanisms that explain the onset and the progression of human diseases. In this regard, the department of human genetics at Leiden University Medical Centre, a Wf4Ever partner, investigates the genetic background and molecular mechanisms behind a number of rare and common diseases. We summarize in what follows the golden exemplars:

- ***Metabolic Syndrome***. This golden exemplar describes the situation where the aim is to mine the relationships between the genotype (genetic code) and the phenotype (disease symptoms). This study involves running *in silico* experiments that are enacted by workflows, but also the analysis of relationships between data used as input to the experiments and the data obtained as a result.
- ***EpiGenetics in Huntington's Disease***. In the second exemplar users investigate the mechanisms leading to HD phenotypes. As for the previous use case, this requires the design and modelling of experiments that combine different types of data sets and analysis tools.
- ***Toxicogenomics – experience from a novice user***. The third use case, aims to interpret the effect of gene transcription factor on the gene expression in the small intestines from wild type and PPARalpha-null mice. The user, who is not familiar with workflow managements systems, attempts to design the experiment that can be used for this study by means of the Taverna workbench. Moreover, the user would like to employ pre-existing workflows stored within the myExperiment repository as component (sub-workflows) within the user's target experiment.

As for the astronomical use cases, the requirements elicited from the above use cases will be reported later on in this document. Nonetheless, as we did in the previous section, we would like to stress that with regard to the activities that are the focus of this document, this use case main requirements are related with the last golden exemplar, where an user faces the situation of needing to find and combine workflows relevant to its work, without the previous knowledge or expertise in the field to be aware of their existence (or even recognize them). '

6.6 User roles, user stories and user requirements

From these golden exemplars, as depicted in the proposed methodology in the section 3 Methodology, we have extracted a set of representative user roles that represent significant user types; and after that we extract users desires expressing them in the form of user stories.

The roles that have been identified are contained in Table 2.

Role	Description
Collaborator	The collaborator is a scientist that is working in a group that uses Wf4Ever collaborative platform. A collaborator takes advantage of the seamless integration with its own working environment in a sharing and ubiquitous platform.
Creator	A creator is a scientist conducting an investigation who wishes to collect together its resources that can then be reused or repurposed. This may be for personal re-use (the

	scientist may not yet wish to *publish*).
Reader	A scientist that is looking for related works, state of the art, in its field of research. The reader skims the titles and abstracts of the publications, sometimes delving into the content of publications that it may be interested in re-use or comparison. Readers are also associated with newcomers to a field that will gain new roles as they become more familiar with the research field and research techniques, scaling to comparators, (re)user, publishers, and evaluators.
Comparator	The researcher that is looking for research work similar to the one she or he is working at present. The comparator wants to know if the work has been already published, or if there is one that that execute similar tasks to those present and needed in the comparator's work. A comparator is more interested in the workflows and not so much in metadata, data, authorship and publications related. A comparator may come from a very different scientific domain, since workflows for statistical tasks in biology may be also very useful for a scientist. If a comparator finds a workflow suiting its specific needs it might take the role of (re)user.
(Re)User	The scientist that has a practical understanding of working with workflows (and its associated elements, such as experiments, datasets, etc). The (re)user knows how to extract and replace modules from one workflow and to insert them into its own. Most of the times the (re)user also takes the role of comparator, other times it is another colleague of the comparator and the comparator just goes right through the selected research material someone has identified. Like the comparator, re(user) is more interested in workflows (and associated data such as datasets, experiments, etc.) and not so much in metadata.
Publisher	The scientist who wants to publish an enhanced publication "beyond the pdf". The publisher might be the main author of the publication, though the real person who undertakes the action of publishing it may be one of the co-authors. The publisher wants his digital experiment to be known, and by extension the work done in its group, among the research community.
Evaluator	The scientist that has enough experience in its field of research to evaluate and score published material. The evaluator can provide comments and suggestions to improve the published work both in a methodological perspective and from technical point of view. The evaluator can evaluate and rate not only the whole scientific material but also specific components relating to quality criteria in reproducibility, repeatability of the results and re-usability/usefulness.

Table 2 Identified users roles.

In terms of these roles, we have defined the following user stories that aim at describing user's actions in a technical agonistic manner, focusing just in their actual needs. We only include those that we believe relevant to the scope of this document, which are the following:

<table>
<tr><th colspan="3">As a Creator ...</th></tr>
<tr><th></th><th>I want to...</th><th>so that ...</th></tr>
<tr><td>UR1.1</td><td>collect scientific data</td><td>I can conduct an investigation</td></tr>
<tr><td rowspan="2">UR1.2</td><td rowspan="2">aggregate existing scientific resources (workflows, datasets, experiments, etc…)</td><td>I can conveniently access related (updated) resources from a single place</td></tr>
<tr><td>I can be sure that I have a matching collection of resources</td></tr>
<tr><td>UR1.3</td><td>reference scientific data stored elsewhere</td><td>I can aggregate data that is larger/more complex/restricted</td></tr>
<tr><td>UR1.4</td><td>describe the relationships between aggregated scientific resources</td><td>other researchers can see how the resources fit together</td></tr>
<tr><td>UR1.5</td><td>be recognised as the creator of an a given scientific</td><td>I get credit</td></tr>
<tr><td>UR1.6</td><td>assign a persistent URL to an aggregate scientific content</td><td>I can include the link in my book</td></tr>
<tr><td rowspan="2">UR1.7</td><td rowspan="2">record which web services were used by workflow</td><td>I can track web service changes</td></tr>
<tr><td>give citations to external resources used</td></tr>
<tr><td rowspan="2">UR1.8</td><td rowspan="2">embed other’s publications</td><td>I can later find related reference material/citations</td></tr>
<tr><td>I can get information when designing my experiment</td></tr>
<tr><td rowspan="2">UR1.9</td><td rowspan="2">record notes while designing workflow</td><td>later pick up my thoughts around a part of workflow</td></tr>
<tr><td>disseminate reasoning behind my design decisions</td></tr>
</table>

Table 3 Creator user story

As a Contributor ...		
	I want to...	**so that ...**
UR2.1	provide new or updated scientific data/results	investigations are up to date
UR2.2	modify scientific contents	I can fix a known error with a workflow or investigation
UR2.3	be credited for my contributions to a research publication	I get credit and make tenure
UR2.4	have access to the work and scientific content carried out by another researcher	I can contribute to shaping this work before it's public

Table 4 Contributor user story

As a Collaborator …		
	I want to…	**so that …**
UR3.1	provide scientific content	It can be incorporated or used in an investigation
		other researchers can review the processing performed; other researchers can repeat the processing performed

Table 5 Collaborator user story

As a Reader ...		
	I want to...	**so that ...**
UR4.1	find relevant scientific materials	I can understand the field
UR4.2	browse an overview	I can determine whether there is something useful for me
UR4.3	survey the field	check whether something has been done before
UR4.4	examine the relationships between resources	I can understand the relationships between resources (and the creators/owners of such resources)

UR4.5	access data	I can look at it and use it for my own purposes
UR4.6	access metadata	I can see where data/methods came from
UR4.7	follow the steps taken in certain research activity	I can understand the investigative process or method
UR4.8	find workflow by purpose	I can investigate different approaches to the same problem
UR4.9	find workflows according to their reputation	I can investigate approaches that have been acknowledged as being correct for the same problem

Table 6 Reader user story

As a Reviewer/Evaluator ...		
	I want to...	**so that ...**
UR5.1	rerun a scientific investigation	I can validate that the results are as given
UR5.2	examine the relationships between research resources	I can validate those relationships
UR5.3	access scientific data	I can validate the data used
UR5.4	check if external scientific data has changed	I can determine if results are still valid
UR5.6	examine the resources related with a given research	I can determine the source of those resources
UR5.7	rate research concept	I can recommend materials to colleagues

Table 7 Reviewer/Evaluator user story

As a Comparator ...		
	I want to ...	**so that ...**
UR6.1	compare some scientific data/results with others	I can determine whether the investigation is novel
		I can understand the differences between investigations
		I can consider reusing it in the future

Table 8 Comparator user story.

As a (Re)User ...		
	I want to ...	**so that ...**
UR7.1	build a new workflow based on an existing one	I can do something new with less effort
UR7.2	build a new workflow based on an existing one	I can use an existing, known, validated methodology
UR7.3	build a workflow using components/parts of another workflow	I don't have to investigate how to use a service
UR7.4	run an existing workflow with new data	I can get new results by using existing procedures
UR7.5	use results from an existing investigation as input to a new one	I can build on existing results
UR7.6	use data from an existing investigation as input to a new one	I can build on existing data
UR7.7	see and navigate versions of a workflow	I can use the latest working version
		I can better understand a workflow by understanding how it has evolved
		I can see how the latest version of a workflow differs from an earlier version I may have used
UR7.8	extract scientific content	I can reuse that content for other investigations
UR7.9	be able to work in a collaborative fashion on the same workflow and associated scientific material	I can work with my teamwork without having to carry out the manual synchronization process based in the interchange exchange mails with custom scripts, workflows and draft versions of papers and having to update
UR7.10	find workflows and their associated scientific material according to their reputation	I can reuse approaches that have been acknowledged by the community as being correct for the same problem
UR7.11	the system to take into account my past workflow and dataset selections	in next iterations with the system it filters scientific material that doesn't match my past criteria

Table 9 (Re)User user story.

As a Publisher ...		
	I want to...	**so that ...**
UR8.1	publish certain scientific result	it is available for others to see or use
UR8.2	provide references to existing scientific data/result	they can be cited (leading to credit)
UR8.3	be able to advertise certain scientific data/result	It reaches its target audience
UR8.4	restrict access to subparts of a research work	comply with license restrictions
		keep data owner happy

Table 10 Publisher user story

User requirements that are informally described in the user stories have been projected in to technical requirements, as described in the methodology, see section 3 Methodology. In the following tables we relate these projected requirements with their originating user requirements. ECM stands for evolution and collaboration model, REC corresponds with the recommendation activity, and finally, SHA corresponds with the sharing activity.

As a Creator ...	
User requirement	**Technical requirement(s)**
collect scientific data	(REC) Discoverable RO, (REC) Repurposeable RO
aggregate existing scientific resources (workflows, datasets, experiments, etc…)	(ECM) Alive RO, (ECM) Referenceable RO, (REC) Repurposeable RO
reference scientific data stored elsewhere	(ECM) Alive RO, (ECM) Referenceable RO
describe the relationships between aggregated scientific resources	(ECM) History RO, (ECM) Comparison
be recognised as the creator of an a given scientific	(ECM) Plagiarism
assign a persistent URL to an aggregate	(ECM) Referenceable RO

scientific content	
record which web services were used by workflow	(ECM) Referenceable RO
embed other's publications	(ECM) Referenceable RO
record notes while designing workflow	(ECM) Reliable RO

Table 11 Creator user requirements projection onto technical requirements

As a Contributor ...	
User requirement	**Technical requirement(s)**
provide new or updated scientific data/results	(ECM) Alive RO
modify scientific contents	(ECM) Roll-back, (ECM) Fix and diagnose
be credited for my contributions to a research publication	(ECM) Referenceable RO, (ECM) History RO, (ECM) Versioning
have access to the work and scientific content carried out by another researcher	(ECM) Referenceable RO

Table 12 Contributor user requirements projection onto technical requirements

As a Collaborator …	
User requirement	**Technical requirement(s)**
provide scientific content	(ECM) Alive RO

Table 13 Collaborator user requirements projection onto technical requirements

As a Reader ...	
User requirement	**Technical requirement(s)**

find relevant scientific materials	(ECM) Quality and trustworthiness, (REC) Discoverable RO
browse an overview	(REC) Discoverable RO (REC) Content-based
survey the field	(REC) Discoverable RO
examine the relationships between resources	(ECM) History RO, (ECM)Reliable RO, (ECM) Comparison, (SHA) Visualization
access data	(ECM) Alive RO
access metadata	(ECM) Alive RO (REC) Discoverable RO, (REC) Repurposeable RO
follow the steps taken in certain research activity	(ECM) Historical RO
find workflow by purpose	(REC) Discoverable RO, (REC) Policy-based recommendation
find workflows according to their reputation	(ECM) Quality and trustworthiness, (REC) Discoverable RO, (REC) Repurposeable RO, (REC) Policy-based

Table 14 Reader user requirements projection onto technical requirements

As a Reviewer/Evaluator ...	
User requirement	**Technical requirement(s)**
rerun a scientific investigation	(ECM) Repeatable RO
examine the relationships between research resources	(ECM) Comparison
access scientific data	(ECM) Alive RO
check if external scientific data has changed	(ECM) Alive RO

examine the resources related with a given research	(SHA) Visualization
rate research concept	(ECM) Quality and trustworthiness, (ECM) Social approach, (REC) User feedback

Table 15 Reviewer/Evaluator user requirement projection onto technical requirements

As a Comparator ...	
User requirement	**Technical requirement**
compare some scientific data/results with others	(ECM) Referenceable RO, (ECM) Comparison

Table 16 Comparator user requirements projection onto technical requirements

As a (Re)User ...	
User requirement	**Technical requirement(s)**
build a new workflow based on an existing one	(ECM) Referenceable RO
build a new workflow based on an existing one	(ECM) Referenceable RO
build a workflow using components/parts of another workflow	(REC) Discoverable RO, (REC) Repurposeable RO
run an existing workflow with new data	(ECM) Repeatable RO
use results from an existing investigation as input to a new one	(REC) Discoverable RO, (REC) Cross-Boundary adaptative
use data from an existing investigation as input to a new one	(ECM) Alive RO, (REC) Discoverable RO, (REC) Repurposeable RO
see and navigate versions of a workflow	(ECM) Alive RO, (ECM) Referenceable RO, (ECM) Versioning
extract scientific content	(ECM) Alive RO

be able to work in a collaborative fashion on the same workflow and associated scientific material	(ECM) Collaborative construction, (ECM) Social approach
find workflows and their associated scientific material according to their reputation	(ECM) Social approach, (REC) Discoverable RO
the system to take into account my past workflow and dataset selections	(REC) User feedback

Table 17 Re(User) user requirements projection onto technical requirements

As a Publisher ...	
User requirement	**Technical requirements(s)**
publish certain scientific result	(ECM) Referenceable RO, (REC) Discoverable RO
provide references to existing scientific data/result	(ECM) Referenceable RO
be able to advertise certain scientific data/result	(REC) Discoverable RO, (REC) Cross-Boundary adaptative RO (REC) Repurposeable
restrict access to subparts of a research work	(SHA) Access and Security

Table 18 Publisher user requirements projection onto technical requirements

7 Technical Requirements

7.1 Collaboration and evolution model

7.1.1 Research Object Dimensions Requirements

Ontology engineering usually starts by analyzing the domain and application requirements [45] . Hence, for the development of an evolution and collaboration model for Research Objects, we have identified a set of requirements, based on the dimensions stated above, for the two aspects of the model, i.e. evolution and collaboration. With respect to the evolution aspect, the following requirements are specified:

- ***Repeatable Research Object requirement.*** The evolution and collaboration model should provide the mechanisms to identify and characterize different versions of a Research Object and the resources it encapsulates. The goal is to enable scientists to access the same version of the services and data sources involved in the experiments described by a Research Object. This will in turn allow the repetition of the experiments, e.g., to verify or validate the results.
- ***Alive Research Object requirement.*** Research Objects are subject to many different changes, especially because the resources they aggregate can be of a very dynamic nature, e.g., databases. As such, the evolution and collaboration model should enable the modelling of the transformations the Research Object goes through its lifetime. In particular, the model should be flexible enough to cope with all the possible changes a Research Object can undergo.
- ***Referenceable Research Object requirement.*** The ability of representing different versions of a Research Object with the evolution and collaboration model, as noted above, will also play an important role supporting references to Research Objects. In particular, the evolution and collaboration model should provide the means for the identification of such versions, which would enable their referencing as the Research Object evolves.
- ***History Research Object requirement.*** The evolution and collaboration model should allow keeping a complete track of the changes that the Research Object undergoes. The information to be modeled includes not only a representation about the changes themselves, but also about how those changes were applied (e.g., chronological order, dependencies between them); and possibly why they were conducted (e.g., argumentation). Moreover, this information should be kept at different of abstractions, from the Research Object level to the level of the individual resources it encapsulates. Besides, the information should be detailed enough in order to support advanced operations with the Research Object, such as rolling back, fix and diagnose errors, comparison of Research Objects, etc.

With respect to the collaboration aspect, the following requirements are specified:

- ***Reliable Research Object requirement.*** An important source of information about the Research Object supporting the collaboration of scientists is the justification of why and how particular decisions were made. Based on such justifications, scientists can better interpret a Research Object, its objective and speed up the collaborative work during the construction and adaptation of a Research Object. Hence, the should enable the representation and tracking of such justifications as the Research Object evolves.

- ***Referenceable Research Object requirement.*** In order to encourage the collaboration of scientists in the construction of Research Objects, it is important to ensure that it is possible to reference those Research Objects. The evolution and collaboration model should provide the means for the identification of Research Objects and of the resources it encapsulates, which will enable referencing and citing not only the Research Object as a whole but also its individual parts.

7.1.2 Projected User requirements

With respect to the evolution and collaboration model, the following technical requirements have been projected:

- ***Versioning requirement.*** It should be possible to keep different versions of Research Objects. In general, two types of Research Object have been identified so far, a live or working Research Object and a published or archived Research Object. The former is expected to change frequently, while the latter is expected to be immutable. Hence, versioning will have to deal mostly with *live* Research Objects. Versions should have to consider changes at different levels, from the Research Object itself, such as its metadata, to the resources it encapsulates. Moreover, it should be possible to do automatic versioning of metadata (contributors, access privileges, modification date, etc.) to trace the provenance of these metadata. As a result, scientists will be able to access concrete versions of Research Objects and/or their associated resources, e.g., workflows, and to keep track of the evolution of those objects resulting from community efforts.
- ***Roll-Back requirement.*** Research Objects can change many times during their lifetime. Hence, scientists working with a Research Object may want to go back to a previous state of the Research Object, for instance, to retrace steps or to discard new changes introduced. For this requirement is not enough just to recover a previous state of the Research Object, but also being able to reproduce the steps (backwards) to move from one state to a previous one.
- ***Fix and diagnose requirement.*** During the lifetime of the Research Object, it could happen that some errors were introduced while making changes. So, in a similar situation as in the previous requirement, it may necessary to go back to a previous state of the Research Object. However, in addition to that, it should be feasible to make a diagnosis and potentially to fix the errors introduced. To this end, a complete log of the changes performed and how they were performed will be required.
- ***Comparison requirement.*** One of motivations of modelling the evolution of Research Objects is to compare different versions of the Research Object, and in general to compare different states of the same Research Object or with similar Research Objects. This can be useful, for example, in aiding scientists to understand how a Research Object has changed from a previous state, to analyze the transformations, or to find individual contributions to the Research Object.

With respect to the collaboration aspect, the following requirements are specified:

- ***Collaborative construction requirement***. Scientists should be able to work collaboratively in the construction of Research Objects. The involved members of the team may include scientists across disciplines with different expertise, and they may play different roles in such a collaborative process. For example, we could identify investigators (doing most of the scientific analysis), supervisors/advisors (the ones giving mostly scientific guidelines and revisions), domain expert advisors and others. Moreover, the collaboration among scientist may involve curation activities, following some predefined editorial process. That is, a process defining the operations and actions the scientists should follow to propose and to approve/reject changes, depending on their role and the state of the Research Object. Such information regarding the roles and process should be part of the collaboration model.
- ***Social approach requirement.*** Current workflow systems have limited collaborative support for sharing and reusing workflows and related resources. In fact they focus only on making workflows available for sharing. A more social approach is required to really improve the collaboration of scientists. A prerequisite for this will be the modelling of the relationships and interactions between scientist and the scientific resources they collaborate with, as well as the community generated information, such as ratings, reviews and usage logs.
- ***Quality and trustworthiness requirement*.** Scientists are only willing to reuse and extend good quality, trustworthy and coherent scientific work. Hence, the collaboration model should allow keeping trace of relevant information about those aspects of Research Objects, such as authors, contributors, applications, process used for its creation, reviews, etc.
- ***Plagiarism requirement*.** In order to encourage the collaboration of scientists, special attention has to be paid to credit assignment and recognition of individual contributions. This includes, on the one hand a model for the characterization and identification of those resources, which would enable their citation and attribution in an easy and straightforward manner. And on the other hand, the detection of plagiarism. For the latter, an evolution and collaboration model would provide also some mechanisms to detect it, for example by using the changes in the Research Object.

7.1.3 Activity specific requirements

Regarding the evolution and collaboration model the activity specific requirements that we have acknowledged so far are:

- ***Research Object model compliance requirement*.** The evolution and collaboration model should be built on top of the Research Object model that is being specified in Wf4Ever. In fact, the evolution and collaboration model should extend this model with additional elements to represent aspects such as changes, versions, collaborative processes, etc. Hence, evolution and collaboration model should be compatible with the Research Object model, both syntactically and semantically.
- ***Accessibility requirement*.** The evolution and collaboration model should be accessible and processable for both humans and machines. The former can be achieved by the usage of natural language concept names, while the latter can be achieved by the usage of Web-compatible representation languages (see low-level requirements).

- ***Usability requirement.*** The evolution and collaboration model should reflect the needs of the majority of scientists, e.g., as reported by Wf4Ever case studies, but at the same time it should allow proprietary extensions and refinements in particular application scenarios (e.g., different scientific domains). Usability can be maximized taking into account multiple metadata types. For instance, based on the NISO (*National Information Standards Organization*) recommendation [35] , we could have structural metadata to represent, e.g., the physical and logical structure of a Research Object, descriptive metadata to represent information about the content of the Research Object, e.g., for discovery and identification purposes, administrative metadata to represent information that will help managing the Research Object, e.g., when and how it was created, technical information, etc.
- ***Interoperability requirement.*** The evolution and collaboration model should be available in a form that facilitates metadata exchange among applications. The syntactical aspects of interoperability can be covered by the usage of standard representation languages (see low-level requirements). Semantic interoperability can be ensured by means of a formal and explicit representation of the meaning of the metadata entities (see next requirement).
- ***Knowledge formalization requirement.*** As noted above, the evolution and collaboration aspects of Research Objects should be formally and explicitly represented. In order to do so, the evolution and collaboration model will be formalized as an ontology. This ontology should reuse knowledge from the Research Object model, and it should be designed modularly to enable the integration of both models; and to ease its extension it in the future as needed. Among others, this ontology should allow the representation of different Research Object versions, of changes at different abstraction levels, including information of how they are applied, of the users and their roles, as well as the collaborative processes followed during the creation of Research Objects.
- ***Methodological requirement.*** The development of the evolution and collaboration model ontology should follow state-of-the-art methodologies for ontology engineering, such as the NeOn[2] methodology [45] , which provide guidelines to ensure that the ontology fulfils the domain requirements, and to reuse existing knowledge as much as possible, thus speeding up the production of the ontology and making it compliant with existing models. Such existing knowledge can be in the form of previous ontologies, or in some non-ontological form, such as thesauri, classification schemes, etc.

7.2 Recommendation requirements

7.2.1 Research Object dimensions requirements

With regard to the personalized Research Object recommendation activity, we identify the following requirements that related with some of the Research Objects dimensions:

[2] http://www.neon-project.org/

- ***Discoverable Research Objects requirement***. The recommender system must provide the necessary mechanisms in order to discover proactively the Research Objects that might be of interest to the user. This activity should be performed considering the following properties:
    - *Safety*. The recommender system should avoid recommendation flooding at all costs.
    - *Completeness*. The recommender system should provide the wider as possible set of relevant Research Object. This property is particularly important since the recommender system will in some cases be used to provide an overview of the state of the art to new scientist.
- ***Cross-Boundary adaptative recommender requirement*.** Recommender systems are inherently vertical and configured to provide recommendations in a single and specific domain. The cross-boundary dimension of Research Objects makes mandatory the provisioning of recommendation on the very same Research Objects following different recommendation heuristics for each scientific community or communities. That implies that the recommender system must be aware of the scientific community that the user belongs to; and secondly it must act in accordance.
- ***Repurposeable Research Objects requirement*.** The recommender system must take into consideration the resources that compose a Research Object and not only Research Objects as a whole. Therefore, when making recommendations to a given user the system might suggest new Research Objects; or just resources that might a useful addition/alternative to the ones already aggregated by the Research Objects that the user is currently using or creating

7.2.2 Projected User requirements

The projected user requirements related with the Research Object recommendation activity are the following:

- ***Reputation requirement*** (e.g. db with data with wrong confused units). The trust assigned the Research Object recommendations generated by the recommender system must take into account reputation. The measure of reputation must be multidimensional; the Research Object must not only rated as a whole but also its constituent resources (e.g. datasets, workflows, provider, etc.) must be also considered.
- ***Content-based requirement*.** The recommender system should provide content-based recommendations based on the way that search and retrieving of scientific content is already performed by researchers, allowing search in fields such as authors, abstract, keywords, publication dates, etc.
- ***User feedback requirement*.** The recommender system must consider user feedback in order to improve its future recommendations.
- ***Policy-based recommendation requirement*.** The criterion for selecting the relevance and suitability of an item (either Research Object or resource) is shared among a group of individuals (researchers of the same scientific field, researchers that belong to a concrete lab, etc.); but is not necessarily valid outside this community. There is a need of means for tailoring specific recommendations in terms of certain policies at least for each use case, and in general for each research community that in the future wishes to make use of the recommender system.

7.2.3 Activity Specific Requirements

- ***Research Objects model aware requirement***. The recommender system must use techniques that exploit the formal models about Research Object that will be developed in the context of Wf4Ever.
- ***Cold Start Requirements.*** Most of the state of the art recommendation techniques rely in historical information. Therefore, the addition of new elements that are neither reflected nor referenced in this background knowledge causes problems. We distinguish between two possible situations:
  - ***New user problem handling requirement***. The recommender system must provide mechanisms to lessen effects of the new user problem. When a new user arrives at the system, there is no sufficient rating information to sketch user's preferences; and there might be also a lack of information about the user itself. Both situations must be tacked by the recommender system.
  - ***New Research Object handling requirement***. Every time a new Research Object is created[3] the recommender system must recommend this new item and make to any of the users of the system that might be interested on it. Unlike the case of the new user problem, the possibility of not having enough information about the Research Object is less probable, since we assume that the information about the Research Object is accessible following the Linked Data principles [1] . Nevertheless, we have a problem regarding the estimation of the Research Object reputation.
- ***Sparse ratings problem handling requirement.*** The sparsity problem typically occurs in systems with large number of items in which there are plenty of items rated only by few users, and many users which rated only few. The set of items rated but just few users would unlikely be recommended, no matter how high its reputation might be. The recommender system should minimize as much as possible this specific situation.
- ***Research Object evolution aware requirement.*** The recommender system must take into consideration as efficiently as possible how the evolution of a Research Object affects to its past and future recommendations. In some cases a change might be considered irrelevant and must not be propagated to the recommendations made on the basis of its previous state; whereas some small changes might result in significant changes in the recommendations about the object. In any case, the recommender system must avoid as much as possible the recalculation of recommendations; it should reuse previously made calculations where possible.

7.2.4 Low-level Requirements

With respect to the Research Object commendation activity, we have identified two low-level requirements:

[3] Note that this requirement applies only to newly created Research Objects; the modification of an already existing one is covered in the research object evolution aware requirement.

- ***Linked data principles compliant requirement.*** Recommendations provided by the recommender system must be accessible using linked data principles (see [1]). Briefly they are:
  - Use URIs as names for entities.
  - Use HTTP URIs so that external agents (either software agents or people) can look up those names.
  - When an external agent looks up a URI, provide useful information, using the standards RDF and SPARQL.
  - Include links to other URIs, so that they can discover more things.

  It is important to understand that recommendations are considered as first class citizens and as such should be treated. In consequence they have their own URI and meta-data associated to them.

7.3 Sharing requirements

7.3.1 Projected User Requirements

The user's requirements related with the sharing activity are the following:

- ***Visualization requirement***. The visualization tool for Collaboration Spheres has to present in a graphical fashion how Research Objects are connected by sharing different resources.
- ***Access and Security requirement.*** Research Objects should have different access and security properties depending on user's role, allowing policies such as making Research Objects (or only part of it) to be freely accessible to all users; permitting a finer-grained definition of access permissions to just a target group of users; and also allowing the definition of partial and temporal access permissions.

7.3.2 Activity specific requirements

- ***Graph-based visualization***. The visualization tool for Collaboration Spheres should provide graph-based visualization. Graph-based visualization offers benefits since expanding nodes and letting the user navigate and explore the content presents information in a single visualization instead of distributed it over multiple screens. In addition, since resources and users might be connected by an arbitrary number of relationships, a proper visualization of these relationships is required reducing the complexity of the graph. Thus, there are some issues to take into account:
  - In order to make it easy for users to see how resources and users are interrelated, the different relation types should be also visualized differently. Then, different kind of relationships should be represented as different kind of labelled edges helping users to comprehend multiple relationships at a time. For example, model-defined relations might be represented as directed or undirected labelled edges depending on whether or not it is a symmetric property.
  - In order to make it easy for users to see how resources and users are interrelated, the visualization tool should offer some simplification mechanisms. Some examples are:

1) Metadata that is only referred by a Research Object is not interested when analyzing relationships with others Research Objects, although that information can be displayed if requested by users.

2) Since the number of crossings edges might exponentially grow with the number of Research Objects, resources, and teams, the number of edges should be reduced as far as possible arranging the graph in a chain. This approach will arrange Research Objects that share the same resources in a chain.

- ***Different visualization for different items***. Understanding the relationships between Research Objects, resources and group of scientists is important in order to understand how research is carried out. However, a powerful and effective visualization form needs to show multiple relationships of different types fitting in different purposes and preventing scientists from information overload. We can differentiate the following relation types that can exist in the visualization:
  - Model-defined relations: relations that have been explicitly defined in the model (e.g. aggregations.
  - Content-based relations: automatically derived relations based on the resources, teams, and people across different Research Objects.
  - Metadata relations (such as author, keywords…)
- **Allow flexible and customizable visualization and user interaction**. As a good engineering principle, the tool should separate the visualization engine and the behaviour assigned to user actions through configuration file. Following this principle will let customize the application to different users and contexts at run-time and defining and implementing new behaviours that can be included in the configuration file at design-time.
- **Navigation on-demand**. Since the size of the available information is unmanageable to be loaded into memory, the visualization tool should allow exploratory navigation based on user actions and running different queries to retrieve the information to be shown.

7.3.3 Low-level requirements

Regarding the sharing activity we have identified a low-level requirement.

- ***Visualization tool for Collaboration Spheres requirements***. The visualization tool for Collaboration Spheres will follow the following technical requirements:
  - A web-based application following current web standards (e.g. HTML 5)
  - It will visualize data, which follows the Linked Data principle of having XML/RDF descriptions and querying SPARQL end-points.
  - It will use the Wf4Ever infrastructure for retrieving information about Research Objects.
  - It will use the Wf4Ever infrastructure for enforcing the access properties for Research Objects.

5. References

[1] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22. Igi Publ. Retrieved from http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jswis.2009081901

[2] Bechhofer, S., De Roure, D., Gamble, M., Goble, C. and Buchan, I. (2010) Research Objects: Towards Exchange and Reuse of Digital Knowledge. In: The Future of the Web for Collaborative Science (FWCS 2010), April 2010, Raleigh, NC, USA.

[3] Bechhofer, S., Ainsworth, J., Bhagat, J., Buchan, I., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Goble, C., Michaelides, D., Missier, P., Owen, S., Newman, D., De Roure, D. and Sufi, S. (2010) Why Linked Data is Not Enough for Scientists. In: Sixth IEEE e--Science conference (e-Science 2010), December 2010, Brisbane, Australia.

[4] Belkin, N. J. and Croft, W. B.: 1992, 'Information Filtering and Information Retrieval: Two Sides of the Same Coin?' Communications of the ACM 35(12), 29-38.

[5] Bonatti, P., Olmedilla, D. Driving and monitoring provisional trust negotiation with metapolicies. Sixth IEEE International Workshop on Policies for Distributed Systems and Networks, pp 14 – 23, 2005.

[6] Bonatti P.A., Duma C., Fuchs N., Nejdl W., Olmedilla D., Peer J., and Shahmehri N. (2006) Semantic Web Policies - A Discussion of Requirements and Research Issues. 3rd European Semantic Web Conference, LNCS 4011, pp. 712-724, Budva, Montenegro, June 2006

[7] Burke R., (2002) *Hybrid Recommender Systems: Survey and Experiments, User Modeling and User-Adapted Interaction*, In User Modeling and User-Adapted Interaction, Vol. 12, No. 4. (1 November 2002), pp. 331-370

[8] Chenu R., Giunchiglia F., Xu H., Birukou A., Maltese E.. (2010) LiquidPub D1.2. Design of the SKO structural model and evolution. Version 2. 2010

[9] Damianou N., Dulay N., Lupu E., and Sloman M.. (2001) The Ponder Policy Specification Language. In Proceedings of POLICY 2001: Workshop on Policies for Distributed Systems and Networks. Lecture Notes on Computer Science, pp. 18-39, 2001.

[10] de Moor A., Leenheer P. D., Meersman R. (2006) DOGMA-MESS: A meaning evolution support system for interorganizational ontology engineering, in: Proc. of the International Conference on Conceptual Structures, (ICCS 2006), Aalborg, Denmark, Springer, 2006.

[11] DRIVER II D4.2 Report on Object Models and Functionalities. Available online at http://wiki.surffoundation.nl/display/standards/Objectmodel+Enhanced+Publications

[12] Flouris G., Plexousakis D. (2005) Handling ontology change: Survey and proposal for a future research direction, Tech. Rep. FORTHICS/TR-362, Institute of Computer Science, FORTH, available at http://www.ics.forth.gr/isl/publications/paperlink/fgeo_TR362.pdf (September 2005).

[13] Goldberg, D. Nichols, D., Oki, B. M., and Terry, D. Using collaborative filtering to weave an information tapestry. Commun. ACM 35, 12 (Dec.1992), 61—70.

[14] Groth, P., Gibson, A., Velterop, J. (2010). The anatomy of a nanopublication. Information Services and Use, 30(1), 51-56.

[15] Guttman, Robert H.: 1998, 'Merchant Differentiation through Integrative Negotiation in Agent-mediated Electronic Commerce'. Master's Thesis, School of Architecture and Planning, Program in Media Arts and Sciences, Massachusetts Institute of Technology.

[16] Haase P., Sure Y., Vrandecic D. (2004) D3.1.1 Ontology Management and Evolution: Survey, Methods and Prototypes, Tech. Rep. D3.1.1, AIFB, University of Karlsruhe; Sekt Deliverable, available at http://www.sekt-project.com/rd/deliverables/wp03/sekt-d-3-1-1-IncrementalOntologyEvolution.V1.pdf (December 2004).

[17] Kagal L., Finin T. W. and Joshi A. (2003) A Policy Language for Pervasive Computing Environment. In Proceedings of POLICY 2003: Workshop on Policies for Distributed Systems and Networks. Lecture Notes on Computer Science, 2003

[18] Klein M., Fensel D., Kiryakov A., Noy N. F., Stuckenschmidt H., (2002) Versioning of distributed

ontologies, Tech. rep., Vrije Universiteit Amsterdam (December 2002).

[19] Klein M., Noy N. (2003) A component-based framework for ontology evolution, in: Proceedings of the IJCAI'03 Workshop: Ontologies and Distributed Systems, Acapulco, Mexico, 2003.

[20] Klein M., Change Management for Distributed Ontologies, Ph.D. thesis, Vrije Universiteit, Amsterdam) (2004).

[21] Krulwich, B.: 1997, Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data. Artificial, Intelligence Magazine 18 (2), 37-45.

[22] Laika's MedLibBlog http://laikaspoetnik.wordpress.com/2010/06/23/will-nano-publications-triplets-replace-the-classic-journal-articles/

[23] Lavoie B. F. Technology Watch Report. The Open Archival Information System Reference Model: Introductory Guide. 2004

[24] Lang, K. (1995) Newsweeder: Learning to filter news, In: Proceedings of the 12th International Conference on Machine Learning, Lake Tahoe, CA, pp. 331-339. Littlestone, N. and Warmuth, M.: 1994, 'The Weighted

[25] Liang Y. (2006) Mini-Thesis: Enabling Active Ontology Change Management within SemanticWeb-based Applications, Ph.D. thesis, University of Southampton (2006).

[26] LiquidPub Project http://project.liquidpub.org/liquid-publications-scientific-publications-meet-the-web

[27] Mons B. and Velterop J., *Nano-publication in the e-science era*, in: Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009), Washington, DC, USA, 2009

[28] Noy N. F., Klein M. C. A. (2003) Tracking complex changes during ontology evolution, in: ISWC-2003 Poster Proceedings, Sanibel Island, Florida, 2003.

[29] Noy N., Chugh A., Liu W., Musen M.,(2006) A framework for ontology evolution in collaborative environments, in: International Semantic Web Conference 2006, pp. 544–558.

[30] Palma, R., Haase, P., Corcho, O., Gómez-Pérez, A., Ji, Q. An Editorial Workflow Approach For Collaborative Ontology Development. In: Proceedings of the 3rd Asian Semantic Web Conference. ASWC 08. Bangkok, Thailand, December 2008. Springer

[31] Palma, R., Haase, P., Corcho, O., Gómez-Pérez, A. (2009) Change Representation For OWL 2 Ontologies". In: Proceedings of the Fifth International Workshop OWL: Experiences and Directions. In ISWC09. October, 2009. Chantilly, Virginia, USA.

[32] Parra Trepowski C. D. Infraestructura web para comunidades y recursos cientificos - Modelo, servicios y métricas

[33] Pazzani, M. J. (1999) A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial, Intelligence Review, 13 (5/6), 393-408.

[34] Piero A. Bonatti, Olmedilla D. (2005) Driving and Monitoring Provisional Trust Negotiation with Metapolicies. Proceedings of the 6th International Workshop on Policies for Distributed Systems and Networks (POLICY 2005), Stockholm, Sweden, 2005.

[35] Press, N. (2004). Understanding Metadata. National Information Standards (p. 20). National Information Standards Organization Press

[36] Resnick P., Varian H. R. (1997) *Recommender systems*. Communications ACM 40, 3 (March 1997), 56-58. DOI=10.1145/245108.245121 http://doi.acm.org/10.1145/245108.245121

[37] Rich, E.: 1979, 'User Modeling via Stereotypes'. Cognitive Science 3, 329-354.

[38] Schafer, J. B., Konstan, J. and Riedl, J.: 1999, 'Recommender Systems in E-Commerce'. In: EC '99: Proceedings of the First ACM Conference on Electronic Commerce, Denver, CO, pp. 158-166

[39] Sebastian A., Noy N. F., Tudorache T., Musen M. A., A generic ontology for collaborative ontology-development workflows, in: Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008), 2008, pp. 318–328.

[40] Schafer, J. B., Konstan, J. and Riedl, J.: 1999, ‘Recommender Systems in E-Commerce’. In: EC ’99: Proceedings of the First ACM Conference on Electronic Commerce, Denver, CO, pp. 158-166.

[41] Schein, A. I., A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In Proc. of the 25th Annual Intl. ACM SIGIR Conf., 2002.

[42] Schmitt, S. and Bergmann, R.: 1999, ‘Applying case-based reasoning technology for product selection and customization in electronic commerce environments.’ 12th Bled Electronic Commerce Conference. Bled, Slovenia, June 7-9, 1999.

[43] Soboroff, I. and C. Nicholas. Combining content and collaboration in text filtering. In 43 IJCAI'99 Workshop: Machine Learning for Information Filtering, August 1999.

[44] Stojanovic L., (2004) Methods and Tools for Ontology Evolution, Ph.D. thesis, University of Karlsruhe (TH), Germany (August 2004).

[45] Suarez-Figueroa M. and Gomez-Perez A. (2008) Neon methodology: Scenarios for building networks of ontologies. In Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management EKAW’08, poster papers, 2008

[46] Tempich C., Ontology engineering and routing in distributed knowledge management applications, Ph.D. thesis, University of Karlsruhe (TH), Germany (2006)

[47] Tonti, G.; Bradshaw, J. M.; Jeffers, R.; Montanari, R.; Suri, N. & Uszok, A. Fensel, D.; Sycara, K. P. & Mylopoulos, (2003) J. (ed.) Semantic Web Languages for Policy Representation and Reasoning: A Comparison of KAoS, Rei, and Ponder. International Semantic Web Conference, Springer, 2003, 2870, 419-43

[48] Towle, B. and Quinn, C.: 2000, ‘Knowledge Based Recommender Systems Using Explicit User Models’. In Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, AAAI Technical Report WS-00-04. pp. 74-77. Menlo Park, CA: AAAI Press.

[49] Ungar, L. H., and D. P. Foster. (1998) Clustering methods for collaborative filtering. InRecommender Systems. Papers from 1998 Workshop. Technical Report WS-98-08. AAAI Press, 1998.

[50] Uszok, A., Bradshaw, J. M., Jeffers, R., Suri, N., Hayes, P. J., Breedy, M. R., Bunch, L., Johson, M.,Kulkarni, S., and Lott, J. (2003) KAoS policy and domain services: Towards a description-logic approach to policy representation, deconfliction, and enforcement. In POLICY 2003. IEEE Computer Society.

[51] Uszok, A., Bradshaw, J.M., Johnson, M.., Jeffers, R., Tate, A., Dalton, J., Aitken, S. (2004) KAoS policy management for semantic Web services. Intelligen Systems, IEEE, pp. 32 – 41, 2004.