Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science

STREP FP7-ICT-2007-6 270192

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

D4.1: Workflow Integrity and Authenticity Maintenance Initial Requirements

Deliverable Co-ordinator: Jun Zhao

Deliverable Co-ordinating Institution: Oxford

Other Authors: Graham Klyne (OXF), Guillermo Álvaro Rey (iSOCO), Jose Manuel Gomez-Perez (iSOCO)

Document Identifier:	Wf4Ever/2011/D4.1/v1.0	Date due:	31/05/2011
Class Deliverable:	Wf4Ever 270192	Submission date:	31/05/2011
Project start date:	December 1, 2010	Version:	v1.12
Project duration:	3 years	State:	Final
		Distribution:	Public

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	

Change Log

Version	Date	Amended by	Changes
1.0	20-05-2011	Guillermo Álvaro Rey	Moved all contents here from the online collaborative-writing tool
1.1	23-05-2011	Graham Klyne	Update section 1.1, 3 and 6
1.2	23-05-2011	Jun Zhao	Update section 1 and 2
1.3	23-05-2011	Graham Klyne	Update section 5 to reflect new wording about technical dimensions, add cross-references, regenerate ToC
1.4	24-05-2011	Jun Zhao	Update and polish section 2
1.5	24-05-2011	Guillermo Álvaro Rey	Updates on s5, s6, correction of typos and references
1.6	24-05-2011	Jun Zhao	Updates on s6, adding more requirement analysis results
1.7	26-05-2011	Jose Manuel Gomez-Perez	Conclusions added, gap analysis and SoA updated.
1.8	26-05-2011	Guillermo Álvaro Rey	Updates on s6, correction of typos and references
1.9	27-05-2011	Jun Zhao	Feedback to points 1-12, 14-17, and 22 of the QA form
1.10	27-05-2011	Graham Klyne	Feedback to points 13 and 18 of the QA form and modified section 5.2.2
1.11	30-05-2011	Guillermo Álvaro Rey	Integration of iSOCO and OXF changes, issues 5, 19, 20, 21, and editorial remarks from the QA, wrapping up.
1.12	31-05-2011	Jose Manuel Gomez-Perez	Final version

Executive Summary

This document presents initial requirements for workflow integrity and authenticity maintenance and evaluation. It applies a systematic methodology to analyze users' requirements for workflow integrity and authenticity and confirm a need for provenance information to support these requirements. It identifies a set of technical provenance requirements, including the different types of provenance information needed, and the technical features required for provenance management and use. The initial list of requirements provides a starting point for our design and implementation work. Requirements will continue to be gathered throughout the rest of the project, and will be revisited and evolved over time, taking into account the agile software development approach we take.

Table of contents

List of Figures

1. Introduction and Motivation

Integrity and authenticity of observation data, applied transformations, and interpretation of the resulting data lie at the heart of scientific research, underpinning the quality of resulting information. Authenticity requires that any data or results presented are exactly what they are claimed to be. Integrity requires that the processes and transformations to which data have been subjected have not introduced any undisclosed distortion or bias or loss in the resulting information. To a large extent, it is necessary to trust the reporter in order to have confidence in the integrity and authenticity of what is reported. But we also look for evidence that supports the conclusions we reach through the exercise of such trust.

In the context of Wf4Ever we propose that an important type of evidence that can support the integrity and authenticity of research results is provenance information. A working definition of provenance by the W3C Provenance Incubator Group is as following: “Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource”. One important purpose of this initial requirements gathering and analysis is to test this proposition about the important role of provenance information in integrity and authenticity assessment, by matching the needs of research users to establish confidence in research results, and projecting these into the technical features and capabilities of provenance information and processing models.

1.1. Technical Context

We make a number of assumptions about the technical environment within which Wf4Ever will be deployed. These assumptions are subject to review, and are expected to co-evolve with the wf4Ever technical architecture. These assumptions are also used to frame some of the technical requirements presented later.

- The system operates in the environment of the World Wide Web, supporting normal Web capabilities of retrieval, linking, etc. As such, URIs are used to denote arbitrary concepts, object types, etc. Concepts and entities manipulated by Wf4Ever are preferably identified using URIs.
- For preference, interfaces between Wf4Ever software components will use HTTP in a RESTful fashion [8] , as this facilitates separation of concerns and interoperability between diverse, independently developed components, and makes extensions easy to integrate with the core system.
- Research data and associated notes and metadata are organized into Research Objects (ROs): Research Objects (ROs), covered in detail in D2.1 [26] , are semantically rich aggregations of resources that bring together data, methods and people in scientific investigations. 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.

  In the context of Wf4Ever we focus on those Research Objects that encapsulate scientific workflows, along with datasets, descriptions and metadata associated with them, and consider how the application of a RO approach can support the preservation of those workflows and the results obtained through their execution.

- A RO contains metadata that allow the type(s) of each component to be identified in some way.
- A RO may contain metadata about provenance of its component elements, and also about itself as a whole.
- There is an expectation that metadata will be represented using RDF, the standard form for data and metadata on the Web. But the underlying data models and concepts should be expressible independently of any particular representation, so that they may be presented using alternative formats as circumstances require.

2. State of the Art

In this section we give a brief overview of related work in integrity and authenticity (I&A) evaluation and provenance research and illustrate how these concepts are rooted in digital preservation systems. Provenance information describes the origin of a resource and the process leading to a particular state of that resource. As reviewed by Artz and Gil [10] , two key motivations for recording provenance information on the Semantic Web are establishing trust of information and evaluating quality of information. Similarly, in Wf4Ever we propose that provenance information should provide one of the key aspects of information for supporting the evaluation of the integrity and authenticity of research results. Our working hypothesis (illustrated by Figure 1) is that provenance information is necessary for information quality and trust evaluation and for I&A assessment and maintenance. Our review of the state-of-the-art will not only provide background knowledge about existing provenance technologies and integrity and authenticity assessment but also show that this important role of provenance information in I&A assessment has not yet been fully exploited by existing work. A knowledge gap must be filled in terms of understanding the association between provenance information and the evaluation of I&A. The requirement analysis result presented in this document takes one first step towards achieving just that.

Figure 1: *Working hypothesis*

2.1. Provenance in Digital Preservation

The Open Archival Information System (OAIS) reference model describes the functional mechanisms (ingestion, storage, and access) involved in a preservation system, specifies a number of user roles, and provides a description of the different types of information packages based on their function in such mechanisms. OAIS and related initiatives like the Open Archive Initiative for Object Reuse and Exchange (OAI-ORE) are introduced in more detail in deliverable D2.1 [26] . Herein we focus on the relation between the different types of information package [3] and their role as containers of provenance information.

The three types of information packages observed by OAIS are amenable to contain provenance information. On the one hand, interactions between system and users evidence relevant sources and types of provenance information and, on the other hand, they are sources of valuable insight for the design of Wf4Ever's research objects, especially as to the encapsulation of provenance metadata. The different perspectives on preserved information objects provided by each information package enable the analysis of the kind of provenance information that is more relevant for each functional mechanism of the preservation system.

The **Submission Information Package (SIP)** is the version of the information package that is transferred from the user to the OAIS when information is ingested into the archive. The SIP contains metadata about the object, which can also include useful provenance information about the producer of the research object, the means used to generate it, dependencies with external resources, including datasets, other objects, and services, time of production, etc. Provenance metadata contained in the SIP can be especially relevant for giving credit to the authors of a particular scientific work comprised by a RO as well as to establish the attribution of that RO, therefore enabling system accountability and transparency.

Upon ingestion, and as an instance of a more general problem, provenance metadata needs to remain interoperable between the representation and data model used by the producer and that of the preservation system. As we will see in the following sections, work on provenance vocabulary mapping and the management of meta-provenance are useful in this respect. This problem can be especially relevant as the number of distributed user communities and the sheer amount of users themselves supported by the Wf4Ever system increases.

The **Archival Information Package (AIP)** is the version of the information package that is actually stored and preserved by the OAIS. The AIP consists of the information that is the focus of preservation, accompanied by a complete set of metadata sufficient to support the OAIS preservation and access services. The AIP is a especially relevant placeholder for provenance information, containing metadata that documents the history of the preserved object, 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 it, and the outcome of these actions. Fixity Information validates the authenticity or integrity of the object, e.g. through check sum, digital signature, or digital watermarking.

The provenance information associated to a RO extends, under the perspective of the AIP, the provenance metadata already captured in the scope of the SIP. While the latter focuses on the provenance generated at the producer's end during the ingestion phase, the former deals with provenance generated during storage and, in general, at any time during the custody of the object by the preservation system.

The **Dissemination Information Package (DIP)** is the version of the information package delivered to the consumer users in response to an access request. The DIP concept emphasizes the fact that the information package disseminated by the OAIS to the consumer may differ in form or content from that residing in the archival store. The DIP is the information package metaphor where the preservation system encapsulates all the data and metadata, including provenance information, which more intimately supports RO reuse.

In combination with the notions provided by the SIP, and especially by the AIP, the DIP supports collaboration, sharing, and reuse amongst communities in a preservation system. From a provenance point of view, the provenance metadata captured, managed, and used as part of both the SIP, AIP, and DIP perspectives are essential in order to have a complete view of the chain of custody of a research object, and consequently for computing its integrity and authenticity throughout time. These information package metaphors will inform the design of ROs in order to allow for provenance metadata to be generated, used, and managed as identified from the technical requirements presented in the previous sections of this document.

2.2. Provenance-related Models and Vocabularies

In this section, we give a brief description of various state-of-the-art provenance data models. Although in Wf4Ever we will present out provenance information using Semantic Web, like RDF, the underlying provenance data model and concepts should be expressed independently of particular technologies. Therefore, our review not only covers state-of-the-art provenance vocabularies/ontologies but also conceptual provenance data models.

A provenance data model is a foundation for any provenance-based applications. A number of provenance models have been created over the years, driven by needs from a wide range of application domains, such as e-Science, knowledge representation and reasoning, digital preservation as well as the Web. Some of the most actively developed state-of-the-art provenance models include, but are not limited to:

- The Proof Markup Language [19] : is a provenance data model stemmed from the knowledge representation community, aiming to capture provenance information during the process of reasoning.
- The Provenance Vocabulary Model [20] : is a provenance data model catered for the needs from the emerging Linked Open Data web, aiming to provide specific terms to capture provenance of data creation and access on the Web.
- Provenir provides the foundation for the Provenir upper ontology [15] , which is designed to be extensible to supply domain-specific concepts to support describing and querying provenance information in specific domains, such as bioinformatics or sensor network.
- The Open Provenance Model [21] is a generic provenance data model driven by many years of community effort and collaboration, aiming to provide generic terminologies to facilitate interoperability between provenance data models from different application domains and systems.

To represent our provenance information using Semantic Web technologies, we need provenance ontologies that allow us to publish this information with RDF. Most of these models have been formally represented as an ontology, using languages like RDFS or OWL. There are some other relevant provenance vocabularies that are only represented as an ontology but not described by a data model. These include, but are not limited to:

- The Web of Trust Schema [http://xmlns.com/wot/0.1/]: provides a vocabulary for describing how the validity of data has been assured through being encrypted or signed, relating encrypted data to its key, keys to their users and so on.
- The SWAN Provenance, Attribution, and Version Vocabulary [http://purl.org/swan/1.2/pav/]: is a lightweight vocabulary for keeping track of data provenance and attribution of authoring in an application environment that is heavily based on integration of data from different sources.
- Semantic Web Publishing Vocabulary [http://www4.wiwiss.fu-berlin.de/bizer/WIQA/swp/SWP-UserManual.pdf]: is “an RDF-Schema vocabulary for expressing information provision related meta-information and for assuring the origin of information with digital signatures”. [22]

- Changeset [http://purl.org/vocab/changeset/]: is a vocabulary for describing changes to RDF-based resource descriptions.
- PREMIS (Preservation Metadata: Implementation Strategies) [http://www.loc.gov/standards/premis/]: provides a data dictionary to describe provenance of archived, digital objects (such as files, bitstreams, aggregations).

This large number of contemporary provenance vocabularies obviously makes it difficult to achieve interoperability between provenance information published by different parties, using different vocabularies. A provenance vocabulary mapping report [1] produced by the W3C Provenance Incubator Group (XG) provides a valuable insight into the relationship between these vocabularies.

2.3. Provenance Vocabulary Mapping

The W3C Provenance Incubator group[1], formed between September 2009 and November 2010, was set up to serve the purpose of analyzing the state-of-the-art requirements for provenance on the Semantic Web and identifying a roadmap in the area of provenance for Semantic Web technologies, development, and possible standardization. Therefore, one of the tasks by the incubator group is identifying and analyzing state-of-the-art provenance vocabularies for Semantic Web and producing a report to guide a possible future standardization.

Provenance representation is still an emerging discipline in its process towards standardization, thus current provenance vocabularies are in general biased towards the particular application domains where they were proposed. The mapping process allowed an alignment among the different vocabularies from different provenance communities in order to extract a minimal set of core terms for provenance representation. It is this core model that is especially interesting for us in order to apply provenance information for IQ & Trust evaluation for I&A maintenance in scientific preservation systems, an unexplored area of application until now. This will provide us with a starting point for provenance representation in this area, allowing us to assess the current state of the art in this direction and to do an appropriate gap analysis

In this mapping task [1] , the incubator group members selected 10 state-of-the-art provenance-related models/vocabularies, all of which introduced in Section 2.1. In order to identify the common ground and gaps of coverage among these models, the task force members decided to choose one of the provenance models as the central model to compare the similarities and differences between these models. The Open Provenance Model was chosen as this central model because: i) it is a generic provenance model that is not targeted at a particular application domain; and ii) it is a community data model that draws upon several years of efforts from representatives of different provenance projects and research groups. The choice of OPM was not random but based upon a clear consensus among the XG group members.

The mapping result shows that largely, the concepts from the 9 provenance models and vocabularies can be mapped nicely to the core terms from the OPM. Some of the vocabularies provide more fine-grained

[1] http://www.w3.org/2005/Incubator/prov/

definitions than the OPM, such as the Provenance Vocabulary, because they were created catering for specific needs from a specific domain. Such vocabularies can be potentially defined as either an OPM profile or an extension to OPM. However, at the same time, some other vocabularies, such as Dublin Core and the SWAN Provenance Authoring and Version Vocabulary, provide a more loosely-defined terminology. Although they are either created for generic needs or aimed at maximum reusability, they have less formal expressivity than OPM and must be reused with care.

To take this mapping result forward, the XG members identify a set of core terms that are expressed in a language-neutral manner. It is this set of terms that is used as the starting point for the W3C provenance working group, which is set up to create standard provenance exchange model and languages. These terms should also be considered as the starting point of the provenance model in Wf4Ever. Briefly, this list of terms includes the following [2] [2]:

1. **Resource**: represents both static and dynamic (mutable or immutable) resources on the Web, It is related to concepts like Artifact from OPM and IdentifiedThing from PML. However, the actual semantics of resource is still subject to further discussions, particularly under the context of the Web architecture.
2. **Process execution**: refers to execution of a computation, workflow, program, service, etc. It is meant to refer to a query. Resource should be used for this purpose.
3. **Recipe link**: provides a standard way to establish a pointer. The concept of recipe is not yet defined.
4. **Agent**: refers to entities (human or otherwise) involved in the process execution. An agent can be the creator or contributor.
5. **Role**: is used to distinguish different functions a resource or agent plays in a process execution.
6. **Location**: is an identifiable geographic place [23] . Typically a location is a physically fixed point, typically on the surface of the Earth, though locations can be relative to other, non-earth centric coordinate reference systems.
7. **Derivation: expresse**s the dependency relationship between resources.
8. **Generation**: expresses the relationship between a resource and the process execution that generated the resource.
9. **Use**: expresses the relationship between a resource and the process execution that used the resource.
10. **Ordering of Processes**: expresses an order relationship between process executions. The semantics of this order is not specified. It can represent just the fact that one execution happened before another in terms of term or one caused/trigger the other one.

[2] Note that in the final XG group report no formal definitions were given for these terms. By the time of writing providing definitions for these terms is an ongoing activity in the W3C Provenance Working group. Definitions given in this document reflect our own understanding about these terms and are subject to changes in the future development.

11. **Version**: is a fuzzy notion for one of two or more similar objects with a derivation history between them.
12. **Participation**: represents entities involved in a process execution. They might not have played as an active role as an agent in the execution.
13. **Control**, it is a subclass of participation. Related to this is a notion of "responsibility", an entity that stands behind the artifact that was produced.
14. **Provenance Container**, is used for encapsulating a set of provenance statements.
15. **Views or Accounts**, represents a description or a point of view by a set of provenance statements from one or more observers.
16. **Time**: is used to record temporal information about the occurrence of a process.
17. **Collections**: is notion for one object being part of another.

It is this list of common terms that is expected to drive the standardization effort of the ongoing W3C provenance working group as well as the provenance representation in Wf4Ever.

2.4. Meta-Provenance

So far our provenance review has been focusing on models and vocabularies capturing one group of provenance-related information, i.e. the origin information about a resource and the process that led to the specific state of that resource. However, there is another group of provenance-related information that has been drawing growing attention in the provenance community, which provide annotation-like meta-statement about a provenance statement or a set of provenance statements. Some examples of such meta-statement could be:

- when someone said that a data product D was created by agent A, i.e. the provenance of the provenance statement itself;
- who made these statements about the process that created data product D, which include the tools used to create D, the script used to guide the creation process, parameter setting of the creation, and etc. This is an example of making a provenance description about a collection of provenance statements.

This type of *meta-provenance* provides an extra level of contextual information about provenance statements, e.g. describing who provided some provenance statements, when, under what circumstances. This extra contextual information is particularly useful when we want to verify the integrity and authenticity of research results based on provenance information. Provenance information such as how a data result was generated can be used to verify whether the data presented is what it is claimed to be. Before we can use this provenance information to perform this verification, we must have some confidence on the integrity and authenticity of provenance information itself. This is particularly important when working with data on the open Semantic Web. Such provenance statements can be made by anyone, at any time, using any tool.

Without any knowledge about the provenance of this provenance information, we risk drawing conclusions about trustworthiness based on totally untrustworthy data.

In Wf4Ever, we represent our provenance information using RDF. However, plain RDF does not provide structure to make meta-statement about RDF data. Some Semantic Web technologies, such as RDF Reification [24] and Named Graphs [25] , provide the structure for representing such meta-statements. However, RDF Reification is known for its vagueness in semantics and awkwardness in querying. Although named graphs are not part of the current RDF core standard, they are widely supported in RDF storage systems (Jena, 4Store, etc.). Further, named graphs can and sometimes are implemented in the web simply by virtue of posting different graphs at different URIs, and the SPARQL standard query language or RDF contains query structures that assume the existence of named graphs. Additionally, the recently started W3C RDF Working Group (http://www.w3.org/2011/rdf-wg/wiki/; started in January 2011) is expected to provide a much more efficient, standardized solution to this pressing issue.

Apart from efforts in the semantic web community, the Metadata Provenance Task force (launched in June 2010) (http://dublincore.org/groups/provenance/) from DCMI (Dublin Core Metadata Initiative) also proposes the metametadata modelling, in which metadata is regarded as data and metametadata is the metadata about metadata. So far this task force is focusing on documenting exemplar use cases (http://wiki.bib.uni-mannheim.de/dc-provenance/doku.php?id=europeana_example) and they are aiming to align their data model and implementation with the recently formed W3C Provenance Working Group (launched in the end of April 2011). The Meta Object Facility from OMG [http://www.omg.org/mof/] also provides an extensible model driven integration framework for defining, manipulating and integrating metadata and data in a platform independent manner.

2.5. Integrity and Authenticity Evaluation

As mentioned in the beginning, in Wf4Ever by **authenticity** we mean the evaluation of whether data or results presented are exactly what they are claimed to be, and by **integrity** we mean the verification that the processes and transformations to which data have been subjected have not introduced any undisclosed distortion or bias or loss in the resulting information. Clearly, the evaluation of I&A should be highly related to the evaluation of information quality. Information quality (IQ) is related to the study of the quality of information, such as its reliability, accuracy, or up-to-dateness. Its assessment is widely interpreted as “an aggregated value of multiple IQ-criteria” [6], such as accuracy, completeness, believability, and timeliness. For the purpose of our I&A study, we need first to have an understanding about a range of IQ-critieria and existing work on IQ assessment in order to achieve an overview of the role of provenance in IQ assessment.

Naumann [6] and Bizer [9] provide a thorough review of a range of IQ-criteria. The following table provides an outline definition of the reviewed IQ-criteria, some of which are closely related to I&A while some not. These definitions provide a foundation for our understanding about information quality and for our conclusion about the association between provenance information and I&A assessment.

IQ dimensions	Definitions
Believability	The extent to which information is regarded as true and credible [4] . In a sense, believability is the expected accuracy. Synonyms: error rate, credibility, trustworthiness
Objectivity	is the degree to which information is unbiased and impartial [4] .
Reputation	is the degree to which the information or its source is in high standing [6]. Synonyms: credibility
Consistency	implies that two or more values do not conflict with each other [5] . Synonyms: integrity, compatibility
Verifiability	is the degree and ease with which the information can be checked for correctness [6] . Synonyms: naturalness, traceability, provability
Accuracy	is the degree of correctness and precision with which information in an information system represents states of the real world [7] . Accuracy is often used synonymously with data quality, as opposed to information quality. For us, data quality or accuracy is only one aspect of the overall information quality Synonyms: data quality, error rate, correctness, reliability, integrity, precision
Completeness	is the degree to which information is not missing [4] . Synonyms: coverage, scope, granularity, comprehensiveness, density, extent
Timeliness	is the degree to which information is up-to-date [14] . Synonyms: up-to-date, freshness, currentness
Relevancy	is the extent to which information is applicable and helpful for the task at hand [4] Synonyms: domain precision, minimum redundancy, applicability, helpfulness
Amount of Data	is the extent to which the volume of data is appropriate for the task at hand [4]. Synonyms: essentialness
Understandability	is the extent to which data is easily comprehended by the information consumer [4]. Synonyms: ease of understanding
Interpretability	is the extent to which information is in appropriate languages, symbols, and units, and the definitions are clear [4]. Synonyms: clarity of definition, simplicity

Accessibility	is the extent to which information is available, or easily and quickly retrievable [4]. Synonym: Availability, retrievability, reliability
Security	is the degree to which data is passed privately from users to the data source and back [6] Synonyms: privacy, access security

The assessment of IQ can be regarded as “the process of assigning numerical values (IQ-scores) to IQ-criteria”. Similarly, the assessment of I&A can also be achieved by an aggregation of the assessment of relevant IQ-criterion. Generally speaking, there are three different approaches for IQ assessment: i) based on inputs from users; ii) based on the content of evaluated information; and iii) based on metadata information about the assessed object. Although we aim for an automatic assessment approach as much as possible in Wf4Ever, we do not preclude a user-based assessment approach.

The **user-based** approach mostly relies on the use of questionnaires to gather input from information consumers. For example, Lee et al. use a questionnaire as the assessment instrument, to measure user feedback for each IQ criterion on a scale of 0 to 10 [11] . Such methods, although being quantitative, are based on subjective, user input. The confidence in these rating-based quality scores is closely dependent on the number of the users participating in the assessment.

The **Content-based** IQ measurement uses information itself for IQ assessment. It can be applied to structured data, semi-structured data, or natural language text [9] . For structured data it is often possible to reliably extract information by queries that can then be used to measure its quality. However, assessing quality of unstructured data relies on an accurate technique to process the information content and automated text analysis can often be inaccurate.

Naumann performed a possible matching between general-purpose **metadata** attributes and the assessment of IQ-criteria [6] . His analysis shows that metadata attributes, such as date, publisher, and contributors, are highly related to the assessment of IQ-criteria including accuracy, believability, reputation, and objectivity, all of which can probably be included in the assessment of authenticity and integrity.

IQ assessment is known to be hard. Despite a large amount work on conceptualizing information quality, much fewer work have proposed concrete methods for quantifying the quality assessment [9] . Provenance metadata, as one of the most important source of information for IQ assessment, is even less fully exploited. Golbeck and Mannes [13] use provenance information about user-contributed annotations on the Web to compute trust values and to recommend how much a user should trust others. However, they do not compute the trustworthiness of the annotations themselves but the users. The Orchestra system [16] uses provenance information about who performs an update and the steps taken during the updates to compute whether one should accept or deny an update request from a peer. However, their trust policy about which data source to trust or to prefer is pre-configured, still relying on input from information consumers. Ballou et al. use provenance information, such as the time when the data was obtained, for measuring the timeliness of the data item [17] . However, their proposal is more applicable in a product management system. A lot of

the concepts in their framework are not applicable to dealing with digital data, such as definitions about life span of a (data) product. Hartig and Zhao [18] propose a methodology for computing trustworthiness of data using provenance on the Web. However, their work is preliminary and requires access to much richer provenance information corpus to further test their hypothesis. This shows a clear gap in existing work for making trust judgments based on provenance information. One of the advantages of Wf4Ever in this respect is that it provides a more regulated scenario than the open Web, hence it should be easier to produce provenance-rich metadata in ROs and to use such metadata for IQ and I&A evaluation.

Artz and Gil [10] have shown that on the Semantic Web the two most widely described motivation for recording provenance information are trust and IQ. Furthermore, based on more than 30 use cases contributed by a variety of sub-communities of Semantic Web, the final report of the aforementioned W3C Provenance Incubator Group [http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/] also concluded that provenance information is similarly expected to play important role in the evaluation of understandability, accountability and trust, in dealing with imperfections and debugging of failures and errors, and in supporting explanation of differences. These reviews and analysis of requirements from a large and diverse user community clearly demonstrate the important role of provenance expected in the evaluation of information quality. In this requirement document, we aim to further proof this point of view based on our systematic requirement analysis.

3. Methodology

The ultimate goal of our requirement analysis is to understand the requirements from concrete application domains to workflow integrity and authenticity, which will then guide us on understanding the need of provenance information for maintaining and evaluating the integrity and authenticity of Research Objects. In this project, we aim to address the I&A maintenance and assessment by mainly making use of provenance information about research objects. Hence, our requirement analysis should drive our understanding about what provenance information is needed and how provenance information should be used for this area of study.

With this goal in mind, we adopt a user-led methodology in order to extract a set of initial requirements for the areas of integrity and authenticity maintenance and the needs for different types of provenance information, starting from a very general statement of user goals.

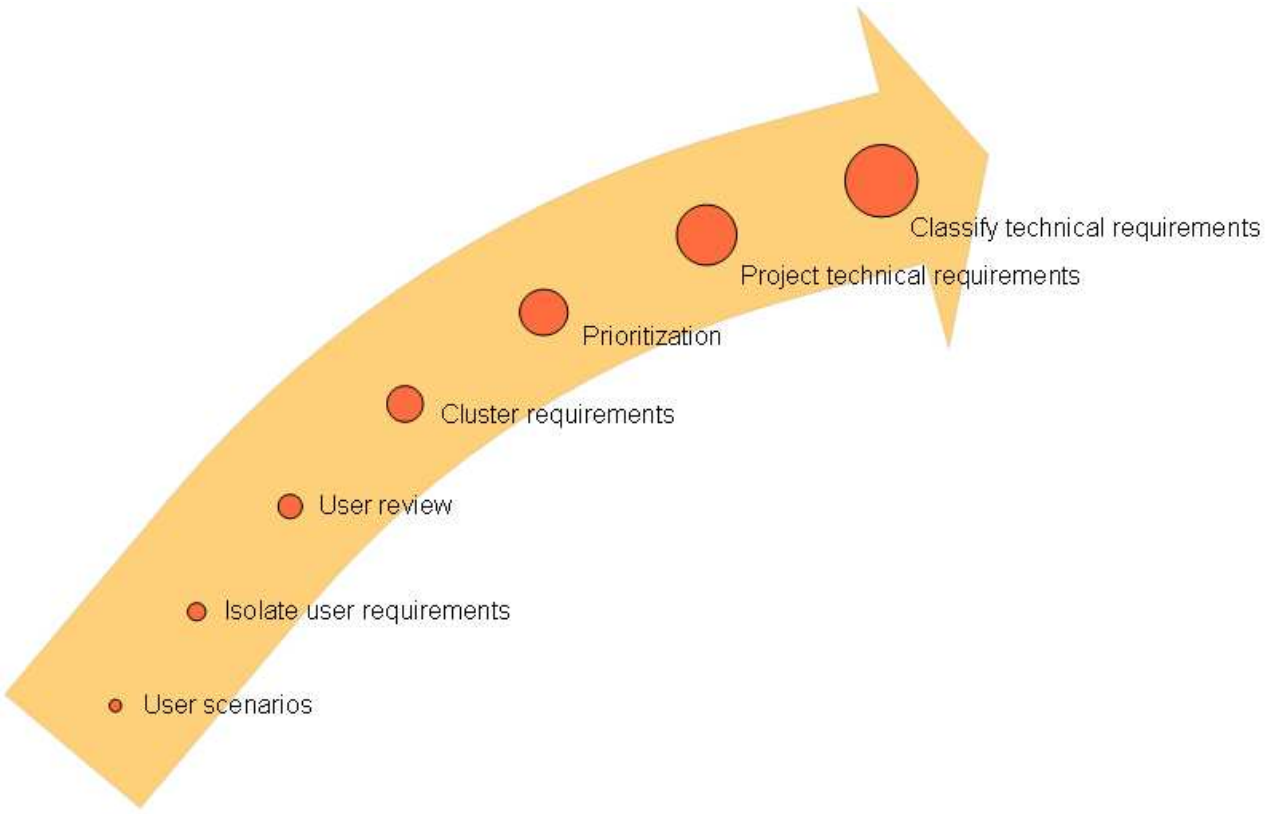

Figure 2: *Outline of the methodology used for extracting initial requirements for the areas of Integrity and Authenticity maintenance. Note that although the methodology is used to gather initial requirements, this process subject to iteration and refinement as the project progresses.*

The methodology, depicted in Figure 2, can be summarized as follows:

- Gather user scenarios: we begin by studying and analyzing some user scenarios or “golden exemplars” from our target users' domains of activity.
- Isolate user requirements: from the scenarios and exemplars, we distil a set of user requirements.
- User review: we review the distilled requirements with the users

- Cluster requirements: requirements are grouped and consolidated as appropriate
- Assess impact and prioritize requirements (optional): user attempts to evaluate requirements in terms of the impact they have on day-to-day work (optional)
- Project technical requirements: extract technical requirements from the user requirements
- Classify technical requirements: organize the technical requirements into different categories (dimensions)

The intent is that the above process will be iterated as required. In the early stages, we do not try to capture all possible requirements, but rather to focus on those areas that the users perceive will have the greatest impact on their day-to-day work. This is done in the full expectation that subsequent development iterations will introduce and analyze new requirements, the details of which may well depend on experience with implementation and user experience with the earlier requirements.

For example, in the bioinformatics domain, we focus initially on a scenario based on a research user of workflows ("reuser"), leaving to future iterations the consideration of scenarios concerning comparison and review of workflows and their outputs. This follows our user partner's clear conviction that the greatest initial impact will be achieved by addressing the workflow users' requirements.

The methodology steps are expanded in the following sections.

3.1. Gather user scenarios

We begin by gathering user scenarios or golden exemplars from our target users' domains of activity.

For example, the astronomy golden exemplars used as starting point deal with:

- Quantities Propagation: updating large amounts of tabular data from external repositories and files through different mathematical equations.
- Sources Extraction: retrieving a list of potential objects fulfilling specific criteria.
- 3D Modelling: 3-dimensional galaxy modelling through external data.

The biological scenario used is based initially around a description of a research workflow user's envisaged interaction with and creation of workflows. (Other user scenarios will be examined later.) The starting point is a prose scenario written by a biological user illustrating typical goals and concerns tackled in the course of their research activity.

At this stage, we ask our users to articulate their goals and requirements in terms that are are germane to their day-to-day work, without particular reference to the technologies we aim to develop. To the extent that the users are also technically knowledgeable, this separation of concerns might not be as clear-cut as idealized here, but the important thing is that the users feel they can focus on what they would like to achieve rather than the technology needed to deliver on those goals.

It is also important that the users first capture what is most important to them, rather than produce a comprehensive list of everything that they might one day find useful. There will be plenty of opportunity later in the overall process to capture additional requirements as they assume greater importance.

3.2. Isolate user requirements

From user-provided materials we distill or infer a set of free-standing 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.

Following a common agile development practice for describing "User Stories" (http://en.wikipedia.org/wiki/User_story), we aim to capture details that can be articulated in the form:

> *"As a (type of user) 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. Further discussion of this template for articulating user requirements can be found in a blog post at http://blog.mountaingoatsoftware.com/advantages-of-the-as-a-user-i-want-user-story-template. The author of this blog post, Mike Kohn, has been credited as being the originator of this template (http://scrummethodology.com/scrum-user-stories/).

3.3. Review user requirements

The distilled requirements are reviewed, initially to ensure that the original scenario has been properly understood and represented.

The user representative is asked to review the distilled requirements, to confirm that they do properly reflect the scenario described, or to clarify any misunderstandings there may be in its interpretation. This review process will typically elicit additional requirements and/or benefits that can be included.

Meetings both via teleconference and face-to-face with subject-matter experts in the astronomy and bioinformatics domains, have been useful in order to trigger and refine the extraction of requirements.

3.4. Cluster requirements

The 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 the face of such redundancy, it can be difficult to prioritize the requirements.

The 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. This process must at least be reviewed by the user, and preferably take place

with their active involvement, to ensure that important nuances of their requirements are not lost or distorted in the process.

It is the result of this consolidation process that are listed below as user requirements.

3.5. Assess impact and prioritize requirements (optional)

(This step is not strictly needed for the requirements gathering and analysis, but any information collected here will later provide useful guidance for setting priorities for technical development work.)

The user is asked to assess the importance, in terms of impact on their day-to-day work, of the various requirements identified. This does not need to be a definitive or complete prioritization of all items, but should aim to highlight those which can have immediate and significant impact on their work. Opportunities will arise throughout the development process to re-assess priorities as experience and understanding of the overall environment develops.

Where provided, user assessments of impact are included with the user requirements listed below. The impact assessment can be assessed in two areas:

- an indication of the extent to which the indicated requirement impacts on the research user's day-to-day work ("benefit to researcher"), and
- an indication of the extent to which the requirement bears on efficiency or effectiveness of the scientific community as a whole ("Importance to Science").

In the analysis presented here, each of these areas has been scored on a range of 1 to 5, where 1 indicates minimal importance, and 5 indicates a maximum impact or benefit associated with addressing the corresponding requirement.

3.6. Project technical requirements

Until this point, the focus has been entirely on articulating and prioritizing user goals and requirements, in principle without reference to the technical mechanisms by which they may be realized. In this step, the user requirements are assessed in the context of a technical deployment environment, and corresponding technical requirements are proposed whereby the requirements can be satisfied.

At this time, assumptions about the nature of the technical environment (stated in terms of the current system technical architecture) should be articulated. These assumptions, in turn, will impose constraints on the evolution of the technical architecture. The technical assumptions applied for the requirements listed here are described in Section 1.1 of this document.

The technical requirements should co-evolve with both the emerging user requirements, and with the developing technical context and architecture. As such, they provide a clear point of linkage between the user requirements and the technical implementation, and provide a basis for justifying technical design decisions, and prioritizing feature development, according to user requirements.

3.7. Classify technical requirements

We associate technical requirements with broad technical areas, which we refer to as *technical dimensions*. We expect to identify common technical areas by this classification of technical requirements, and to guide us towards appropriate areas of further technical work. Classification of requirements into dimensions was originally proposed as a starting point for the requirements gathering but we found that the technical nature of the dimensions (to the extent that they are based on assumptions about the technical nature of the system) means that it is often not meaningful to apply such classification until the technical requirements have been identified.

Thus, in this methodology, the organization of requirements into dimensions occurs as the final stage of the requirements gathering and documentation process. The process allows for the possibility that the co-evolution of technical requirements with both user requirements and technical architecture may result in requirement dimensions that were not envisaged at the outset.

We reuse technical dimensions proposed by prior work. We reuse the provenance dimensions proposed by the incubator group to identify i) common *provenance content* that is required by a technical requirement; ii) common provenance management features; and iii) a set of common provenance use types. Because integrity and authenticity are two important aspects of the more general dimensions associated with information quality, we also reuse the information quality dimensions from earlier work in the information quality community (see Section 2.1) to identify common technical features required for assessing or maintaining an IQ dimension, such as verifiability, believability, or consistency.

The process of identifying and associating technical dimensions with requirements involves a judgement call by the researcher, and as such should be subject to review and revision. In any case, it is entirely expected that subsequent development iterations will bring revised perspectives to bear, resulting in revision of the requirements and analysis.

4. Use Cases Summary

In this section, we cover the characteristics of the use cases we take as base for our requirement-extraction methodology, coming from the Astronomy and the Bioinformatics domain. We also describe the roles that have been identified as relevant for our study.

4.1. Summary of the Astronomy Use Case

4.1.1. Roles

Different user roles have been identified for the Astronomy domain in D5.1 [28] . These are briefly described below, though there is more information on them in the aforementioned document. It is worth noting that we have mainly focused on three particular user roles (marked in the listing below between brackets), namely the “Comparator”, the “Reuser” and the “Evaluator”, for they were the most relevant ones in the context of the scenarios under study in our area of research; other roles might be considered later for next iterations of the methodology.

- **Collaborator:** The *collaborator* is working in a group which uses Wf4Ever collaborative platform and tools, taking advantage of the seamless integration of his own working environment in a sharing and ubiquitous platform.
- **Reader**: The *reader* is looking for related works, state of the art, in his field of research, skimming the titles, keywords, themes and abstracts of the published research objects.
- **[Comparator]**: The *comparator* is looking for research objects similar to those where he is working at present, mainly to know if the work he is doing has been already published as a research object.
- **[Reuser]**: The *reuser* knows how to work with scientific workflows, and how to extract and replace modules from one workflow and insert them into his own; most of the times he has also taken the role of comparator.
- **Publisher**: The *publisher* wants his work and his group to be known among the community, being the main author of the published research object.
- **[Evaluator]**: The *evaluator* has enough experience in his field of research to evaluate and score a published research object, and he can provide comments and suggestions to improve the methodology showed in the research object from a scientific but also technical point of view.

4.1.2. Scenarios

Deliverable 5.1 covers three use cases that have been identified as “golden exemplars”, which are representative of the experiments that are performed within the astronomy field. These are briefly described here:

- **Propagation of quantities**: This scenario 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 a workflow is enacted every time values of variables, on which the magnitude, distance or intrinsic luminosities depend, are updated.

- **Extraction of galaxy samples**: In this scenario, the astronomer aims to 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.
- **Modeling of 3D data of galaxies**: This scenario showcases the need for processing and transferring large volume of data, which are 3D binary cubes with two spatial dimensions and a third one associated with the velocity of the gas emitting the light captured. Such data are generated and processed using workflows.

Through the analysis of these golden exemplars, we have extracted a set of user and technical requirements, as we will report later in the document. It is worth noting that, with respect to the area under study in this deliverable, we will be paying special attention to the need of version methods and access mechanisms to the data related to the workflows.

4.2. Bioinformatics Use Cases

4.2.1. Roles

Three bioinformatics researcher roles have so far been identified.

- **(Re)user** - research user of workflows: A scientist who is looking for workflows as a basis for answering her research question. This could be the person doing the experiment or her supervisor (usually together). The goal is to find a workflow (or several) that forms the basis of one that can be used to address the problem at hand.
- **Comparator**: A researcher who compares their work, or planned work, with that of other researchers. This is typically done before a new experiment (hence it is part of a reuser's activities), but also at later stages (e.g. before publishing). The goal is to know what other researchers are doing, not necessarily to use their workflows as the basis for new work.
- **Reviewer**: The reviewer scenario will be based on a researcher who needs to evaluate the work of a peer. It has similarity with comparator role, but it is (in principle) independent of the reviewer's own work. The goal is to evaluate the quality of a peer's work.

4.2.2. Scenarios

We have focused on the research user of workflows, whom we expect to achieve the greatest impact of workflow support. The following scenario is based on the research activities of roles and use cases described in the *Genomics Workflow Preservation Requirements* D6.1 [29] . For the purpose of requirements extraction and analysis, we have focused on a specific scenario provided by our bioinformatics research partner, which is described below.

(Re)user - research user of workflows

Dennis is a researcher is looking for workflows that will help him interpret data from a gene expression experiment. In this scenario student Dennis has made a conceptual workflow that takes the result of a gene expression experiment (activity values of all genes under two conditions: with/without a chemical compound). The wet laboratory experiment was done by others then Dennis. He makes a note of the origin (including a paper reference). The initial hypothesis is that the chemical compound disturbs gene expression. It is yet unknown which genes and what biological processes are affected. The conceptual workflow first performs one of the standard data pre-processing steps for the type of data Dennis has (Affymetrix gene expression array), then it uses a statistical test to filter those genes that are significantly differentially expressed between the two conditions, and finally it performs an enrichment test to find those pathways that are most prominent among the filtered genes. The latter requires an annotation process, where each gene is coupled to the pathways it was once implied in other experiments (there is a database for that: KEGG).

Dennis is new to workflows, so he wishes to start with an existing workflow. For each component he will search myExperiment for keywords. He then wishes to understand the workflows: look into them, perform test runs with test data and his own data, and see other peoples logs. When he finds workflows he does not understand, Dennis is inclined to create his own workflow with his own scripts. He will receive scripts from colleagues and perform tests that his colleagues are familiar with. As such, he can learn what his workflow is doing. This will help him interpret his results.

Ultimately, the workflow may suggest for instance that the set of differentially expressed genes has the Wnt pathway as most common denominator. This pathway is well known for embryogenesis and cancer, information he finds on the internet. He makes a note of that. It will lead to the hypothesis that the chemical compound, may have effects on embryogenesis and/or cancer. This is now his interpretation of his experiment that he wishes to link to his experiment and the processed data. Dennis notes that in a next cycle he will want to perform another workflow that specifically tests this hypothesis, rather than perform an enrichment test. He will then look for a workflow that performs a 'global test', and replace this part in his workflow with the global test workflow. In his log he indicates this fact. In this case he will link the result of this test (most likely a new hypothesis) to the previous experiment and in particular to the initial hypothesis. At some point, he wishes to be able to retrieve this past information and the interrelationships among his hypotheses.

Assuming his finding and new hypothesis are valuable and new, he will publish his results. The publication has cleaned information, sufficient for evaluating his hypothesis and rerunning the one workflow and the one dataset that lead to this result.

5. Requirements for Integrity & Authenticity Maintenance

The primary goals addressed by the requirements articulated here are to understand the needs for enhancing the integrity (or soundness) of scientific research outputs, and also their authenticity (that they properly represent what they are claimed to represent), from the researchers' perspective. To this end, we have asked our astronomy and bioinformatics partners to articulate how they expect to use workflows, and the capabilities that they consider as important for establishing integrity and authenticity of their results. In the process of developing these requirements, we make specific reference to provenance-related requirements, as articulated by prior and ongoing work to develop a provenance framework in the belief (supported by the state-of-art survey summarized above) that provenance information will be one of the key elements upon which information integrity and authenticity assessments are based.

Following the methodology explained in Section 3, we have gathered requirements from use cases in two different domains, namely i) astronomy, and ii) bioinformatics. For both sets of requirements, we took the same approach, distilling in the first place requirements from the user perspective (user requirements labelled UA* for the astronomy domain, and UB* for the bioinformatics one), and mapping those to technical requirements that relate, in this case, to the subject of provenance, and in particular to the topic of integrity and authenticity maintenance (technical requirements labelled TA* for the astronomy domain, and TB* for the bioinformatics one).

In the subsections 5.1 and 5.2 below, we list and justify the requirements extracted per exemplar under study.

5.1. Requirements in the Astronomy Domain

The requirements that we describe in this subsection have been distilled from the astronomy golden exemplars that are briefly addressed in Section 4.1 and extensively covered in deliverable of work package 5 D5.1 [28] .

5.1.1. User requirements

From the three golden exemplars described above, and from a user (and not technical) perspective, we have extracted a list of requirements, which is the base of the following table. As explained in the astronomy role description section, we focus on three particular non-exclusive roles for our analysis, namely i) comparator, ii) reuser, and ii) evaluator. It is important to note, though, that subtle differences in terms of benefits for the researcher have been pointed out in some occasions depending on the particular role of the user. Hence, as a comparator/reuser/evaluator of astronomy workflows,

User Req.	I want to ...	so that I can...	Benefit to researcher	Importance to Science	User comments
UA1	know details about the data used as input to a workflow execution	know under which conditions was the original data retrieved, and the	4: comparator, reuser 5: evaluator	5	It's extremely relevant to know the provenance/ conditions of your data in a sense these two factors are related with quality of the science involved

		provenance of that data itself			in the Wf.
UA2	know the details about the data transformations that take place during the workflow execution	check all the transitory data	4-5	5	For astronomers this is quite importance since they are looking to see what happens when one upgrades properties of objects in catalogues, and how this updates propagates to other quantities.
UA3	know the status of the decision points in the execution of a workflow	see the branches taken in a workflow execution and the rationale behind them	3: comparator 4-5: reuser, evaluator	2	OK as long as it does not affect the quality of the science
UA4	move (back and forth) from different versions of the same RO/workflow	inspect their differences	4-5	4	Very relevant, and the different versions of the same Wf may lead to better quality of science
UA5	navigate relevant metadata information by following the links among them and to have an advanced visualization of this information	they can have a more human-oriented presentation of and access to the metadata information	5	2-3	Really important for the users since we are talking about visual influence over the user independently of his/her role; having all the data available with the touch of a click is a fundamental idea
UA6	visually see the evolution of their ROs/Wfs over time	see how they change from version to version	5	5	A picture is worth a 1000 words, and this is something that is valuable for Astronomy
UA7	express their (subjective) opinion about the quality of external data sources	so that they or their colleagues can use these external data sources with care	3	4-5	This is very controversial matter, as assessing the quality of data/articles/software is quite subjective. (Users might have biased opinions about data/services that make this a very complicated matter.) However, overall the importance for science is huge, as good quality data / reliable services all improve science.
UA8	know the Quality	so that they can	3	4-5	(See previous comment.)

	of Service of external data sources	choose the most reliable ones among services of similar functionality			
UA9	reproduce a scientific experiment (consistency)	obtain the same results if the inputs and transformations are the same	5	5	Very important; it means that if Wf is clear enough to be reproducible and consistent, it saves you time, and it is trustworthy and it also mean that if you fully understand it you can improve it. If methodologies become easily repeatable, the science results can be easily verified by independent teams leading to accuracy and quality.
UA10	debug a workflow	fix it if the results of their execution seem to be inconsistent, incomplete, etc.	5	5	Very important, as this directly affects the quality of you final product

5.1.2. Technical requirements

The technical requirements, to some extent subjective, that can be distilled from the user requirements above are summarised in the following table. The purpose of the *Dimensions* column is to associate technical requirements with broad technical areas, as described in the methodology (section 3.7).

Tech. Req.	User requirement	Technical requirement	Dimensions
TA1.1	To know details about the data used as input to a workflow	Provenance information about the data used as input, which is also properly (semantically) linked with provenance information about ROs and Wfs	*Accessibility, Verifiability*
TA2.1	To know the details about the data transformations that take place during the workflow execution	Information about data transformations and transitory data generated during a workflow execution needs to be accessible	*Accessibility, Verifiability*
TA3.1	To know the status of the decision points in the execution of a workflow	Information about each path taken during the workflow execution and the conditions present at the decision points needs to be accessible	*Understandability*
TA4.1	To move (back and forth) from different versions of the	Metadata needs to permit versioning, with versions semantically linked	*Verifiability, Believability*

	same RO/workflow		
TA5.1	To navigate relevant metadata information by following the links among them and to have an advanced visualization of this information	Tools need to use the links in the metadata to permit Web navigation through the provenance information	*Accessibility*
TA5.2		Tools need to display provenance information as graphs, allowing navigation through that kind of visualization	*Accessibility, Understandability*
TA6.1	To visually see the evolution of their ROs/Wfs over time	Tools need to include a display of the status of a RO/Wf where the evolution over time (versions) are shown by handling a control	*Accessibility, Understandability, Verifiability*
TA7.1	To express their (subjective) opinion about the quality of external data sources	Tools should allow researchers to add their own assessment of external data sources used (both as free form text, and also using computer-processable vocabularies where appropriate).	*Reputation, Accuracy*
TA8.1	To know the Quality of Service of external data sources	Metadata information about the behaviour of the external data sources in terms of response time, availability, etc. Tools that estimate the QoS for users and display such information along with the data sources	*Reputation, Believability*
TA9.1	To be able to reproduce a scientific experiment.	Metadata information needs to contain all the necessary information in order to repeat an experiment and obtain the same results, and tools supporting scientists to use this metadata information to reproduce their experiment	*Completeness, Verifiability, Believability, Consistency*
TA10.1	To be able to debug a workflow	Metadata should record information about previous executions, and differences from the current execution, and tools using this information to explain any inconsistency or incompleteness of previous execution	*Verifiability, Believability*
TA10.2		Intermediate results and trace information should be captured and recorded during a workflow execution, and made available along with the final results.	*Accessibility, Verifiability, Believability*

5.2. Requirements in the Bioinformatics Domain

5.2.1. Scenario: (Re)user - Research User of Workflows - User Requirements

Scenario based on a researcher who is looking for workflows that will help him interpret data from a gene expression experiment. The requirements listed here have been articulated, clustered and assessed as described in the methodology section.

As a research user of workflows:

	I want to ...	so that I can...	Benefit to res-earcher	Import-ance to Science
UB1	discover existing workflows or parts of workflows that do something similar to what is required for my experiment	get ideas for my own work	4	4
		compare with my work ('related work')	2	4
		reuse in my own analysis	5	4
		understand a method	3	3
		reuse the results	5	4
UB2	reference the experiment that provides the input to the workflow (including paper reference, if any)	acknowledge data providers	2	5
UB3	record my initial experimental hypothesis	link the experimental design to its purpose (onset)	2	4
UB4	assemble conceptual workflows	discuss my plans with supervisors/peers and start searching for components	3	4
UB5	adapt an existing workflow to my needs	avoid having to learn (or remember?) everything about creating workflows (note: newbie scenario)	4	4
UB6	search for workflows by keywords about purpose, context, and outcome of experiment	find workflows relevant to my experiment	5	5
UB7	run an existing workflow with different data	understand the workflow, get more biological results efficiently	5	4
UB8	compare results of workflow run with other results	understand the results in biological terms, compare with competition	3	4

UB9	create new workflows with my own scripts	'get on with it' without trying to understand other people's work	5	4
UB10	run workflow scripts provided by colleagues	perform experiments of which the methods are familiar to my direct colleagues, to shorten the start-up phase and lower the risk of unexpected bottlenecks	5	3
UB11	record notes relating to experiment (design and run-time log)	document my considerations during design and execution, linked to design, run, and input/output data.	2	5
UB12	record revised experimental hypothesis	link experimental design and execution to the purpose of the experiment	4	5
UB13	search for 'follow-up' workflows	create a new workflow or revise my own to test the improved hypothesis or new questions derived from the previous results	4	3
UB14	record reasons for workflow revision	Link new workflow to previous workflow	2	4
UB15	link test results and interpretation to initial hypothesis	retrieve past information about and interrelations between hypotheses	4	5

(Derived from. http://www.wf4ever-project.org/wiki/display/docs/Biological+User+Requirements)

5.2.2. Scenario: (Re)user - Research User of Workflows - Technical Requirements

These technical requirements are distilled from the Biological User Requirements listed above, as described in the methodology (Section 3.6), and expressed in terms of the assumed technical context described above (Section 1.1).

A full list of technical requirements identified is in the project wiki[3]. The table below includes only those technical requirements that were judged to be related to dimensions of information quality or provenance, and is a subset of the full table.

The purpose of the *Dimensions* column is to associate technical requirements with broad technical areas, as described in the methodology (section 3.7).

The technical requirements and technical dimensions listed here are to some degree subjective, and as such are should be expected to evolve through wider review and analysis of the user requirements, and also on the basis of ongoing implementation experience.

[3] www.wf4ever-project.org/wiki

	User requirement	Technical requirement	Dimensions
TB2.1	reference the experiment that provides the input to the workflow (including paper reference, if any)	refer to experiments and datasets that are defined by other researchers or projects, hence incorporate external work by reference. Such references provide acknowledgement of work used and indications of the source of included elements.	*Verifiability*
TB2.2		ROs and their main component resources are identifiable using URIs, allowing for incorporation (re-use) of resources by reference rather than copying, and avoiding divergence between multiple instances of a resource.	*Consistency*
TB4.1	to assemble conceptual workflow from common processing elements	Workflow can reference externally defined processing elements and workflows, from which they are constructed	*Verifiability*
TB4.2		One or more mechanisms for describing workflows composed of component process elements, incorporated by reference rather than copying, allowing for re-use avoiding divergence between multiple instances.	*Consistency*
TB5.2	adapt an existing workflow to my needs	Represent relationship between the adapted workflow and original workflow, and indicating the origin of components of the new workflow.	*Verifiability*
TB7.1	run an existing workflow with different data	Mechanisms for executing a workflow with specified input data, allowing workflow results to be re-evaluated.	*Verifiability*, *Believability*
TB7.2		Capability to retrieve and present input data used to run the workflow, and corresponding data from previous runs, allowing review and confirmation of steps taken to yield results.	*Verifiability*, *Believability*
TB8.1	compare results of workflow run with other results	Keep results from multiple workflow executions, along with references to associated inputs, parameters, etc., allowing conditions of execution to be reviewed.	*Verifiability*, *Believability*
TB8.2		Facilities for examining workflow input and output datasets	*Verifiability*
TB8.3		Facilities for comparing workflow input and output datasets	*Verifiability*
TB8.4		providing researchers with information that allows them to understand differences between the outputs of workflow runs	*Verifiability*, *Believability*
TB11.1	record notes relating to experiment (design, run-time log and interpretation)	Mechanism for adding free-form textual description of an experiment, with links to external sources (which may be generally known and trusted within a community) that have been used.	*Verifiability*, *Believability*, *Reputation*

TB11.4		Link interpretation of result to experimental run, so computed basis for interpretation is connected.	*Verifiability*, *Believability*
TB12.1	record revised experimental hypothesis	A defined component type for revised hypothesis, allowing reviewers to understand alternatives considered in reaching a final hypothesis.	*Believability*
TB12.3		Representation of revision relationship between different revisions of an experiment hypothesis.	*Believability*
TB13.2	search for 'follow-up' workflows (as basis for a new or revised workflow to test the improved hypothesis or new questions derived from the previous results)	Record association between different versions of workflows or workflows used to test different versions of a hypothesis	*Verifiability*, *Believability*
TB14.1	record reasons for workflow revision	A defined component type for workflow annotation, linked to a workflow	*Believability*
TB15.1	link test results and interpretation to initial hypothesis	Mechanism to link execution results and associated interpretation to initial hypothesis, allowing reviewers to follow an evolving understanding of the data and results generated.	*Verifiability*, *Believability*

(Derived from: http://www.wf4ever-project.org/wiki/display/docs/Biological+Technical+Requirements)

6. Requirements and Gap Analysis

In this section, we address the requirements captured in the section above, in terms of their implication with respect to the area of Provenance (Section 6.1), relating the requirements to provenance vocabularies (Section 6.2), and finally identifying the key information and quality dimensions that stem from our analysis (Section 6.3).

6.1. Requirements and Provenance

Our technical requirements enable us to identify the needs for provenance information in order to support the identified users' requirements. To organize these needs we inherently use the provenance dimensions[4] used by the W3C Provenance Incubator group to classify requirements for provenance gathered from a wide variety of application domains. Our initial analysis shows that our provenance needs do not all fit into the provenance dimensions defined by the incubator group. For example, our technical requirement analysis shows that we do not need provenance information that describes how a knowledge reasoning system produced an answer to a reasoning result. Therefore, we chose to use only the top three dimensions, i.e. provenance **content**, **management**, and **use**, to classify our provenance requirements.

6.1.1. Content Dimension

The *content* dimension aims to capture the different types of provenance information, i.e. the structure and attributes that would need to be represented in order to support the users' requirements we identified. We conclude that we would need the following different types of provenance information:

- Provenance of the input data to an experiment, such as the original experiment that created the data, or the external data source from which the data was retrieved (TA1.1, TB2.1)
- Provenance of RO and workflows, such as from which source data they are derived (TA1.1, TB4.1, TB5.2)
- Provenance about data transformation steps or processing elements (TA2.1, TA9.1, TA10.1, and TA10.2, TB4.1)
- Intermediate data results and their relationships to the intermediate steps (TA2.1 and TA10.2)
- Decision points in a workflow execution (TA3.1)
- Execution branches caused by user's decisions (TA3.1)
- Versions of ROs, workflows and hypotheses, and relationships between different versions (TA4.1, TB8.1, TB12.3)
- Behaviour of external data sources, such as their performance, reliability over time (TA8.1)
- Provenance about the computational basis for an interpretation of experimental results (TB11.4)

[4] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Dimensions

- Reference to the initial hypothesis that prompted creation of a set of results (TB15.1)

The incubator group concludes that the following types of provenance content would be needed: establish the artifact or **object** that provenance statements are about, establish **attribution** to an artifact (such as the source or entity contributing to the artifact), record the **processes** (steps) leading to an artifact, deal with **evolution and versioning**, provide **justifications** to back up decisions, and capture assumptions or knowledge used to **entail** an inferred result. Our analysis result largely concurs with the conclusion from the incubator group, except for the need for entailment content.

Note that the notion of versioning requirement is also reflected in the requirement analysis of work package 3 (see D3.1 [27]), and addressing challenges related to object versioning is one of the key goals of WP3 (Evolution, sharing and collaboration). We expect WP3 to provide an actual versioning mechanism for research objects, to describe, for example, version numbers of objects etc; while in our provenance work package, we focus more on the version relationship between objects, such as one object being a previous version of another one. We regard this as a more specific type of derivation relationship, which is key in provenance; while providing complete versioning support is out of the scope of provenance solution.

6.1.2. Management Dimension

The *management* dimension aims to clarify mechanisms that make provenance available and accessible in a system. Based on our current technical requirement analysis, we identify a need for the following different types of provenance management mechanisms:

- Identification: we need proper identification for all entities whose provenance information are being describe or who are mentioned in provenance records (TB 2.2)
- Reference/Links: we need proper links and references between these entities described in or by provenance records (TA5.1, TB2.2)
- Accessible: provenance information should be accessible (TA1.1, TA2.1, TA10.2) for navigation and visualization (TA5.1), and for searching for relevant workflows, hypothesis or ROs (TB7.2)

Under the management dimension the incubator group identified the needs for provenance **publication**, making provenance **accessible**, maintaining a **dissemination control** of provenance information, and dealing with **scalability**. We do also have the needs for making provenance information accessible, however, at the current stage of our requirement analysis, the concerns for properly identifying entities and creating links between them are more important. We do envision the needs for dissemination control and deal with scalability would appear in future iterations of requirement analysis and technical designs.

6.1.3. Use Dimension

The *use* dimension captures a diversity of scenarios under which provenance information can be consumed and made useful to users. Our analysis shows some concrete use scenarios, some of which are directly related to our goal of supporting the assessment of I&A.

- Navigate provenance information (TA5.1)
- Visualize provenance as graphs (TA5.1)

- Visualize evolution of ROs/WFs (TA6.1)
- Evaluate Quality of service using past execution information (TA8.1, TB7.1, TB8.1)
- Reproducibility (TA9.1)
- Debug inconsistent/incomplete workflow results (TA10.1, TB8.1, TB8.4)
- Verifiability
    - Provide cross-references between research objects and their components to realize citation by reference (TB2.1, TB4.1)
    - Provide cross-references between the versions of the same research objects (TA4.1)
    - Support review of original conditions of an execution (TB7.1, TB8.1)
    - Facilitate an understanding about experiment result differences (TA10.1, TB8.2, TB8.3, TB8.4)
    - Support annotations to experiments, which express user's interpretations or hypothesis of an experiment (TB11.1, TB11.4, TB15.1)
- Believability
    - Mechanism for executing a workflow with specified input data so that results can be re-evaluated (TA9.1, TB7.1, TB7.2)
    - Support review of experiment results (TB8.1, TB8.2)
    - Support review of experiments by user's annotations and their associated hypothesis etc (TB11.1. TB11.4, TB14.1)
    - Support review of experiment hypothesis evolution (TB12.1, TB12.2, TB13.2)

Our results highlight a set of distinctive provenance use dimensions compared to the one identified by the incubator group, which proposes a list of uses including making provenance information **understandable**, maintaining **interoperability** among provenance information, enabling **comparison** between artifacts, establishing **accountability** and **trust**, handing **imperfection**, and debugging failures. Broadly speaking, our use dimension includes some more concrete cases than the one proposed by the incubator group, such as the need for navigation and visualization can be regarded as an aspect of making provenance information understandable. This gives us a more concrete starting point throughout the agile software development cycle, starting from something simple and easiest to achieve, from moving to tackle more complex and composite tasks.

6.2. Provenance Vocabulary Gap Analysis

In Section 2.3 we present the list of common provenance terms reported by the W3C provenance incubator group's provenance vocabulary mapping activity. These terms are not only used by the current W3C provenance working group as the starting point for standardization but also considered as a starting point for provenance representation in Wf4Ever. Our requirement analysis identifies a list of common types of

provenance information required to support our technical requirements. The following table shows how the list of identified provenance terms can support our identified provenance content requirement.

Provenance Term	Required provenance content	Comments
Resource	Yes	Need to be aligned with Wf4Ever definitions of research objects
Process execution	Yes	Likely to be used in Wf4Ever with the same semantics, i.e. referring to execution of a computation
Recipe link	Yes	To refer to entities like workflow designs or even experiment hypothesis
Agent	Not explicitly	Implicitly required by some provenance content, such as recording user's decision point
Role	No evidence for requirement	
Location	No evidence for requirement	
Derivation	Yes	Likely to be used in Wf4Ever with the same semantics, i.e. dependency relationship between resources
Generation	Yes	Likely to be used in Wf4Ever with the same semantics, i.e. expressing the relationship between a resource and the process execution that generated the resource
Use	Yes	Likely to be used in Wf4Ever with the same semantics, i.e. expressing the relationship between a resource and the process execution that used the resource
Ordering of process	No evidence for requirement	
Version	Yes	Need to be aligned with Wf4Ever definitions of research objects
Participation	No evidence for requirement	
Control	No evidence for requirement	
Provenance container	No evidence for requirement	
Views and accounts	Not explicitly	Implicitly required for support provenance navigation & visualization, etc
Time	No evidence for requirement	
Collections	Yes	Need to be aligned with Wf4Ever definitions of research objects

Our analysis result shows that not all the common provenance terms are required for our current list of identified requirements. This does not preclude that we will not need them in the future. As our requirement gathering processes during the development of the project, we definitely expect more terms will be required for our Wf4Ever provenance model, such as time information.

Some of the common provenance terms can be almost directly used in Wf4Ever without modification of their semantics, such as process execution, generation or use; while others must be more finely tuned and aligned with the technical context of Wf4Ever, particularly our definitions and modelling of research objects.

Although the common provenance terms can support most of the provenance content identified by our requirement analysis, we are still in an early stage to claim that they can provide a complete support for this content. For example, although decision points in a workflow execution can be regarded as a kind of Process Execution, they might require extra features, such as annotations to reflect justifications of such decision points. These provenance terms will eventually be formally defined as an ontology. The actual semantics associated with each term and structure of their relationships might cause reconsideration of the scope and coverage of each term. Our analysis result shown in the above table represents a positional understanding, based on the requirement analysis presented in this document. It can be used as a starting point for our following-on technical design and implementation, and it must be iteratively reviewed and updated to reflect the progress of our understanding about our users and about I&A study.

6.3. Requirements Analysis

Following the Methodology described in Section 3, we have isolated a number of user requirements and ensuing technical requirements. As one would expect, a user-led requirements gathering exercise uncovers a wide range of requirements, many of which are not directly related to information quality. While all of these requirements have been captured and recorded in the project wiki, just those we judge to be IQ related are presented here.

Based on this analysis of technical requirements projected from initial user requirements, it appears that the key information quality dimensions are **verifiability** (ability for another to confirm the results), **believability** (availability of evidence to support the results presented, **consistency** (implying that two or more values do not conflict with each other), **reputation** (which in this context is rather similar to believability) and **accessibility** (the extent to which information is available, or easily and quickly retrievable). This suggests that the next stage of our work on integrity and authenticity should focus in these identified information quality dimensions.

7. Conclusions

This document presents a principled approach for extracting user and technical requirements for integrity and authenticity evaluation in Wf4Ever. Through the application of this method in the Astronomy and Biology application domains, we have gathered meaningful, real-life user requirements and have translated them into actual technical requirements, which will inform the design and development of the Wf4Ever I&A evaluation components. Our main work hypothesis, based on the relevance of provenance-intensive information quality and trustworthiness evaluation for integrity and authenticity maintenance, has been proved for the Wf4Ever scenario. The resulting requirements have been classified along IQ dimensions strongly supported by provenance technologies and our requirements analysis has identified a set of distinctive provenance dimensions that refine and extend those originally proposed by the W3C provenance incubator group. Moreover, the analysis of the coverage provided by current provenance vocabularies identifies a meaningful presence of provenance information types as well as further work on provenance vocabularies in order to address the gathered requirements.

7.1. Future Work

These conclusions provide some initial guidance for technical direction of work related to information quality and provenance, and are by no means final. The current set of user and technical requirements will keep evolving during the lifetime of the project and across iterations including the different stages of development and evolution. We will gather user feedback from early software demonstrations and deployment, which will, no doubt, prompt the users to refine and develop their requirements as they see more clearly how the system can interact with their day-to-day work.

Interaction with users will continue in order to gather a broader and deeper view of their requirements, especially through prototypes and demonstrators based on the requirements that allow for further progress. We will also collect and analyze scenarios from other kinds of biological research users (comparators and reviewers have so far been identified), to expose requirements arising from other parts of the research cycle.

The user domains represented in Wf4Ever (WP5 and WP6) serve as an excellent source of requirements from a large amount of different and complementary perspectives. However, we should not stop at these user communities only. The Wf4Ever platform, and in particular the Integrity and Authenticity maintenance component based on Information Quality and Provenance, aims at facilitating preservation and reuse of scientific knowledge in experimental disciplines, where workflows are a cornerstone.

Towards the design and development phase, we shall engage in deeper discussions with the technical architecture team (WP1), progress on Research Object and workflow management (WP2), and work on workflow evolution, sharing and collaboration (WP3). This will support the exploration of the co-evolution of the technical architecture and expression of the technical requirements in terms of the architectural framework, including component interaction.

8. References

[1] Provenance Vocabulary Mappings W3C Provenance Inbubator Group http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings. Published on August, 2010. Last accessed: May 2011

[2] Provenance XG Final Report. W3C Provenance Incubator Group. http://www.w3.org/2005/Incubator/prov/XGR-prov/. Published on December 2010. Last accessed: May 2011.

[3] Lavoie, B. The Open Archival Information System Reference Model: Introductory Guide. Microform & Imaging Review. Volume 33, Issue 2, Pages 68–81, ISSN (Print) 0949-5770, DOI: 10.1515/MFIR.2004.68, Spring 2004

[4] Leo Pipino, Yang Lee, and Richard Wang. Data Quality Assessment. Communications of the ACM, 45:211–218, 2002.

[5] Massimo Mecella, Monica Scannapieco, Antonino Virgillito, Roberto Baldoni, Tiziana Catarci, and Carlo Batini. Managing Data Quality in Cooperative Information Systems. In Proceedings of the Confederated International Conferences DOA, CoopIS and ODBASE, pages 486–502, 2002.

[6] Felix Naumann. Quality-Driven Query Answering for Integrated Information Systems. Ph.D. Thesis. Springer, Berlin Heidelberg New York, 2002.

[7] Yair Wand and Richard Wang. Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM, 39(11):86–95, 1996.

[8] L. Richardson, S. Ruby: RESTful Web Services. O'Reilly Media, Inc. May 2007.

[9] Chritian Bizer, Quality-Driven Information Filtering in the Context of Web-Based Information Systems, Ph.D. Thesis, Freie Universität Berlin. VDM Verlag, 2007

[10] Donovan Artz and Yolanda Gil. A survey of trust in computer science and the Semantic Web. Journal of Web Semantics: Science, Services and Agents on the World Wide Web. 5(2), pp 58-71. June 2007.

[11] Y.W. Lee, D.M. Strong, B.K. Kahn, R.Y. Wang, AIMQ: A Methodology for Information Quality Assessment, Information & Management 40 (2002).

[12] E. Knorr, R. Ng, V. Tucakov, Distance-based Outliers: Algorithms and Applications, International Journal on Very Large Data Bases 8 (2000) 237–253.

[13] J. Golbeck, A. Mannes, Using Trust and Provenance for Content Filtering on the Semantic Web, in: Proc. of the Models of Trust for the Web Workshop at WWW. Provenance-based Validation of e-Science Experiments, in: Proc. of ISWC.

[14] Beverly Kahn, Diane Strong, and Richard Wang. Information Qualtiy Benchmarks: Product and Service Performance. Commmunications of the ACM, 45(4):184–192, 2002.

[15] S. Sahoo, R.S. Barga, A.P. Sheth, K. Thirunarayan, P. Hitzler: PrOM: A Semantic Web Framework for Provenance Management in Science. Kno.e.sis Center Technical Report, Wright State University (2009)

[16] Z. Ives, T.J. Green, G. Karvounarakis, N.E. Taylor, V. Tannen, P.P. Talukdar, M. Jacob, F. Pereira, The Orchestra Collaborative Data Sharing System, ACM SIGMOD Record 37 (2008) 26–32.

[17] D. Ballou, R. Wang, H. Pazer, G.K. Tayi, Modeling Information Manufacturing Systems to Determine Information Product Quality, Management Science 44 (1998).

[18] Olaf Hartig and Jun Zhao. Using web data provenance for quality assessment. In International Workshop on Semantic Web and Provenance Management, Washington D.C., USA, 2009.

[19] Paulo Pinheiro da Silva, Deborah L. McGuinness and Richard Fikes. A Proof Markup Language for Semantic Web Services. In Information Systems 31 (2006)

[20] Olaf Hartig and Jun Zhao. Publishing and Consuming Provenance Metadata on the Web of Linked Data. In Proceedings of The Third International Provenance and Annotation Workshop (2010).

[21] Luc Moreau, Ben Clifford, Juliana Freire, Yolanda Gil, Paul Groth, Joe Futrelle, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Yogesh Simmhan, Eric Stephan and Jan Van den Bussche. The Open Provenance Model - Core Specification (v1.1), 2009.

[22] Christian Bizer and Richard Cyganiak. Quality-Driven Information Filtering using the {WIQA} Policy Framework In Journal of Web Semantics 7, pp1-10 (2009).

[23] ISO 19112: Geographic information - Spatial referencing by geographic identifiers. 2009-06-17

[24] G. Klyne and J.J. Carroll. Resource description framework (RDF): Concepts and abstract syntax (2004). http://www.w3.org/TR/rdf-concepts/

[25] J.J. Carroll and C. Bizer and P. Hayes and P. Stickler. Named graphs, provenance and trust. In Proceedings of the 14th international conference on World Wide Web, pp613-622. (2005)

[26] Sean Bechhofer, Stian Soiland-Reyes and Khalid Belhajjame. Workflow Lifecycle Management Initial Requirements. Project Deliverable D2.1. EU FP7 Wf4Ever project, May 2011.

[27] Rafael González-Cabero, Raul Palma, Carlos Ruiz and Khalid Belhajjame. Workflow Evolution, Sharing and Collaboration Initial Requirements. Project Deliverable D3.1. EU FP7 Wf4Ever project, May 2011.

[28] Lourdes Verdes-Montenegro, Jose Enrique Ruiz, Antonio Portas, Juan de Dios Santander Vela. Astronomy Workflow Preservation Requirements. Project Deliverable D5.1. EU FP7 Wf4Ever project, May 2011.

[29] Marco Roos, Kristina Hettne, Peter Henneman, Eleni Mina and Dennis Leenheer. Genomics Workflow Preservation Requirements. Project Deliverable D6.1. EU FP7 Wf4Ever project, May 2011.