Enhancing and Abstracting Scientific Workflow Provenance for Data Publishing

Pinar Alper, Khalid Belhajjame, Carole A. Goble
School of Computer Science,
University of Manchester
Manchester, UK
first_name.last_name@cs.man.ac.uk

Pinar Karagoz
Dept. of Computer Engineering,
Middle East Technical University
Ankara, TURKEY
karagoz@ceng.metu.edu.tr

ABSTRACT

Many scientists are using workflows to systematically design and run computational experiments. Once the workflow is executed, the scientist may want to publish the dataset generated as a result, to be, e.g., reused by other scientists as input to their experiments. In doing so, the scientist needs to curate such dataset by specifying metadata information that describes it, e.g. its derivation history, origins and ownership. To assist the scientist in this task, we explore in this paper the use of provenance traces collected by workflow management systems when enacting workflows. Specifically, we identify the shortcomings of such *raw* provenance traces in supporting the data publishing task, and propose an approach whereby *distilled*, yet more informative, provenance traces that are fit for the data publishing task can be derived.

Keywords

Provenance, Data Publishing, Scientific Workflows

1. INTRODUCTION

Computing has recently transformed the practice of science in a fundamental manner. Scientists are increasingly embracing the "Fourth Paradigm" [15]: they are using computational resources to explore large datasets available in community repositories in order to empower new findings. Re-using, combining, aggregating and analyzing datasets using computational tools and analyses have become a commonplace activity. In particular, a large number of scientists are using workflows, as the tool of choice, to systematically design and run computational (a.k.a. *in-silico*) experiments [10].

Using a workflow, an experiment is defined as a network of analysis operations, which can be supplied by third party distributed software programs, e.g., web services, or local programs and scripts. The analysis operations that constitute the experiment are weaved together using data dependencies specifying how the data generated by given analysis operations is used to feed the execution of other operations within the workflow.

Once the workflow is executed, the scientist may want to publish the dataset generated as a result. Such dataset can be used as evidence that supports the hypothesis investigated by the scientist, confirms a known fact, or suggests a new scientific finding. To ensure that such datasets can be re-used by the community and preserved for future analyses and computations, Open Archival Information System (OAIS) Reference Model identifies three kinds of metadata information that need to accompany the published datasets [20]: 1) **Reference** information to unambiguously identify a dataset 2) **Provenance** information that specifies datasets derivation history, origins and ownership, and 3) **Context** information that outlines the dataset's relationships to other datasets and to its environment such as its citations, its dependencies and assumptions. Such metadata information is typically specified manually by a curator, who examines the technical report or paper that describes the dataset to annotate it before its publication. This task can be tedious, time-consuming and error-prone. There is, therefore, a need for a means to assist curators in the data publishing task.

Fortunately, there is a source of information that can be harvested for specifying "Provenance" and "Context" information of datasets generated using workflows, namely the provenance traces generated as a result of workflow execution. Indeed, scientific workflow systems can be easily instrumented to collect provenance regarding the runs of workflows and the lineage of results generated in each run. Given that the majority of scientific workflow systems provide (built-in or plug-in) capabilities for collecting provenance traces of workflow runs [10], we would expect that such provenance traces are being harvested by curators to derive "Provenance" and "Context" metadata for data publishing. That is, unfortunately, not the case. While workflow specifications and the datasets generated as a result of their executions are published by scientists, provenance traces are not being reported in data publishing, instead they hardly-ever leave the personal desktop of the scientist.

Our analysis showed that the main reason is due to the *raw* nature of provenance traces of workflow executions, which make them unsuitable for the purpose of data publishing. In this position paper, we pinpoint the shortcomings of raw provenance traces (Section 2). We argue that a distilled form of provenance is needed to assist data publishers (Section 3). We outline a solution that we propose for generating provenance distillations, by enhancing and abstracting raw workflow provenance, and discuss the challenges that need to be addressed for its realization (in Section 4).

2. SCIENTIFIC WORKFLOWS AND PROVENANCE

In this section, we introduce scientific workflows, present a running example, and identify the shortcomings of raw execution provenance in informing the task of data publishing.

2.1 Scientific Workflows

In recent years, workflow systems [8] have become popular in

EDBT/ICDT '13, Mar 18-22 2013, Genoa, Italy

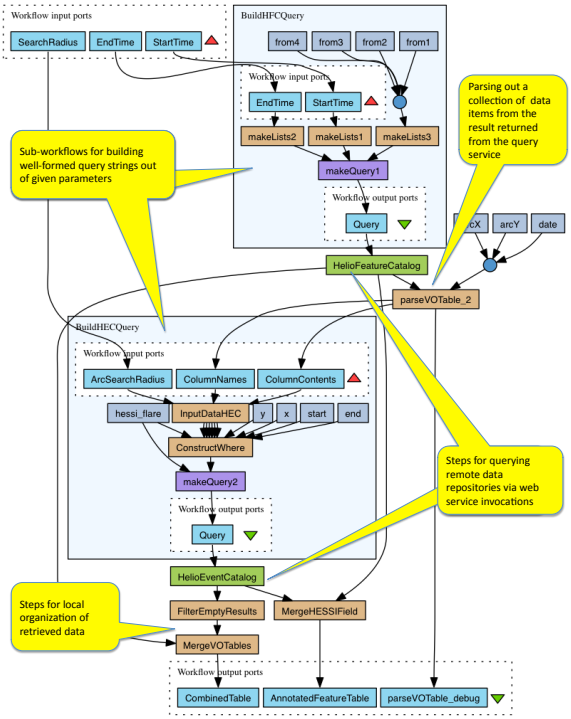

Figure 1: Example workflow from Heliophysics domain.

many scientific disciplines for task automation and in-silico experimentation. Scientific workflows are generally data-driven: they can be thought of as graphs in which the nodes are analysis operations and the edges are data flow dependencies between the constituent operations. Figure 1 illustrates an example of a scientific workflow from the helio-physics domain, which focuses on the study of the "highly energetic events that originate from the sun and propagate through the solar system." [4]. Such a workflow is a data chaining pipeline that gathers and merges data from multiple repositories. The workflow operates as follows: It starts by building up a query to be submitted to the "Helio Feature Catalog" (HFC) to retrieve the active regions detected on the solar surface within a specified time frame. Once the query is executed the result string is parsed into a collection structure and each record is further used as a parameter for generating a query to be submitted to the "Helio Events Catalog" (HEC). The event catalog is queried to retrieve solar flares observed within the locality of the active regions identified in the previous query. This flare data is then aggregated and appended to the Active Region dataset as a new "flare count" column in the MergeHessiField step. The workflow also joins the HFC and non-empty HEC result tables to obtain an overall view of flare events and their regional location.

2.2 Limitations of Raw Provenance Traces

In order to publish the datasets that are produced as a result of execution of the workflow in Figure 1, the provenance traces captured during the execution appears, at first, to be a rich source of information to fulfill the required "Provenance" and "Context" metadata needs of the data publishing task. Unfortunately, such raw provenance traces are not fit for use during publishing due to their following characteristics:

Provenance Obfuscation: Scientific Workflows can be complex with many scientifically significant steps such as analysis, visualizations and data retrieval combined with many "mundane" or insignificant steps such as format transformers, data movers, adapter steps needed for the incorporation of 3rd party resources into the flow [16][14]. Because of this, provenance traces captured during execution are typically a voluminous piece of information containing thoroughly and indiscriminately collected traces for each activity, its input and output data items. In particular, such traces do not specify whether a data item is a significant artifact such as a result of visualization, or whether it is merely a side-product such as an execution status message, or a temporary file address.

Lineage Opacity: The activities that constitute workflows are often opaque black boxes, in the sense that they do not reveal the relationship between the inputs and the outputs. For example, we are unable to determine if a given data item is a query responsible for the retrieval of a data set, or an output is a cleaned revision of the input data of an activity.

Coarse Granularity: As well as the opaque link among data items, workflow provenance report all of the outputs to be dependent on all of the input items. In many scientific disciplines some form of structure to the data typically the tabular form exists[1][2][3]. Moreover, among the activities that occur within a workflow a significant majority of them data organization activities such as filtering, joining and union of collections of data items that essentially relay some (all) of the inputs to the outputs [14]. In all established workflow provenance frameworks, however, due to the workflow system being blind to the structure of data and the nature of the activity the outputs are viewed as freshly generated piece of information computed using all of the inputs. Therefore one item in a collection of result items cannot be traced back to its origin records or it is not possible to identify the activities that are responsible for the initial appearance of a data artifact in a provenance log.

2.3 Related Work

A handful of researchers made proposals with view to address the above shortcomings. The overwhelming volume and the deep traces within provenance has been tackled by providing query languages [23] query-based browsers [3]. In [6] authors describe an approach enabling users to subjectively pinpoint interesting modules in a workflow and partition the workflow into views in order to reduce the amount of response data to provenance queries. As pointed out earlier, workflows contain several mundane operations, and there is little subjectivity in the need to eliminate them during an abstraction.

Domain specific human annotation of workflows, and consecutively, provenance logs, is another approach for making provenance more informative. Typically the domain types of input-output ports or the domain specific function of the activity is specified [19],[21] with annotations. These approaches allow scientists to pose provenance queries using domain terms, however it does not provide an abstraction of provenance. Moreover annotated lineage traces only indirectly hint at the origins of data (e.g. we would not be able to tell which parts of the AnnotatedFeatureTable came from but could only tell that data from both the HFC and HEC service invocations which produce VOTable type outputs are on it lineage path).

Coarse grained provenance has been tackled by adopting approaches in tracking provenance of database queries [7] to workflows. In [2] authors support, "white" box workflow data processing activities and trace which input data records (tuples) give way to which output tuple and through what kind of operations (i.e. join , union). The fine-grained tracing of existential dependency, a.k.a.

[1] http://www.ivoa.net/Documents/VOTable/
[2] http://dataup.cdlib.org/
[3] http://tools.gbif.org/dwca-assistant/

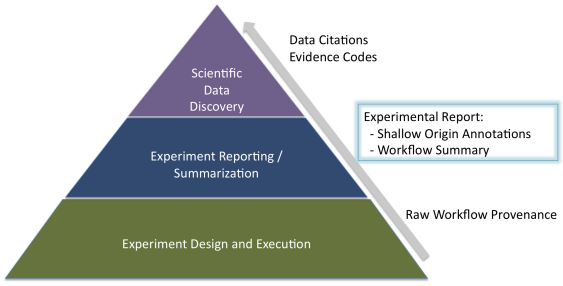

Figure 2: Provenance Pyramid depicting the spectrum of provenance information for different zones of data-intensive science.

why provenance, enables what-if analysis on the relations of input and output data records, however it is blind to value-copying operations occurring in white-box steps. Similarly [17] tracks why provenance for workflows in a non-materialized logical manner with a little more precision (i.e. existential dependencies among attributes of records rather than whole records). We think value-copying relations a.k.a. where provenance for workflows is essential in a data publishing scenario, where curators would inquire about the origin of result artifacts.

All of the these efforts are focused on supporting activities that occur within the local *in-silico* experimental zone during workflow design and execution time. Activities such as workflow debugging, auditing or tracing final and intermediary data items. Existing approaches are not targeted towards data publishing, with the exception of recent work in [11], where authors tackle the issues on integrity-preserving customization of provenance graphs with user-specified data abstraction, anonymization and elimination directives prior to publication. We observe that more research needs to be performed targeted at the middle-layer of the provenance pyramid, where tracking of origins of and context of data and experiments is required, and a small, distilled, yet informative, subset of workflow provenance information is to be generated.

2.4 Plethora of Provenance in Data Intensive Science

We observe that raw workflow provenance and the provenance requirements for data publishing stand at the two ends of a spectrum, which we have depicted as the Provenance Pyramid as illustrated in Figure 2. The Provenance Pyramid takes inspiration from the Data Pyramid [12], which observes that the amount of data that is of value for preservation is inversely proportional to the number of stakeholders interested in the data. During data's staging from a local zone (with a handful of stakeholders) to a community zone (with large numbers of stakeholders), only the significant data items are promoted to the next level.

Therefore, we place raw workflow provenance at the bottom of the provenance pyramid as it contains indiscriminately collected information about every activity and data item within a workflow provenance log. This form of provenance is useful for scientists directly involved in the workflow-based experiment for local execution-time activities such as debugging and steering. At the top we have the provenance information that is of community value. These are small nuggets of information typically specified by manual curation, such as "Evidence Codes" in biological databases[4] or data citations [1]. Community-level provenance is manifested as high-level indicators regarding the derivation method of data, or its origins. This form of provenance is further exploited in the calculation of data quality and trust metrics [13] to assist scientific data discovery.

We argue that a middle-layer of fit-for-publishing [9] provenance information is needed. The middle layer, can be based on a distillation of raw workflow provenance and can be used to inform the top-layer, e.g., it has application areas in Data Citation [1] or Data Usage Metering [18]. As we shall detail in the next section, we propose Distilled Provenance in the form of 1) succinct origin-annotations on result data artifacts and 2) highlights of significant activities in a workflow.

3. PROVENANCE DISTILLATIONS

Provenance information should be fit for the assisting task (or queries) that the user wants to perform. For example, if the user wants to debug a workflow, then one would expect the provenance to be used to answers queries such as "Why is X in the result?" or "What if Y was not in the input would I still get Z?". If, on the other hand, the provenance is used for reporting the experiment and deciding which resources to acknowledge and cite, one would expect to have brief answers to queries such as "What happens in this workflow?" or "Where do these results come from?". We refer to these answers as Distilled Provenance. Distillations should contain:

Shallow Annotations on the results of the workflow as to their data origins and scope. Given the complex nature of workflows, raw workflow provenance contains abundant number intermediate data items that occur on the derivation path of the results. Scientists do not publish raw provenance traces due to the overwhelming depth and complexity of data traces, rendering them unfit. Instead they only share overall workflow results together with manual annotations or citations regarding their origins. Consecutively we argue that the distilled form of origin information could be in the form of **annotations that are minted and attached** to data items during activities such as data retrieval or analysis. These annotations could specify 1) **origin** information e.g. denoting that a VO Table originates from HFC Service deployed at a particular endpoint or 2) **scope** information outlining any significant workflow input parameter that has played a role in the data's generation such as a query criteria, or significant configuration parameter. One important requirement is to be able to propagate origin annotations all the way to the overall workflow results.

A partially ordered set of significant activities within the workflow i.e. the Workflow Summary. To illustrate what we mean by summary, Figure 3 provides the abstraction of the heliophysics pipeline (see Figure 1), containing only the scientifically significant, data minting steps and the significant workflow input and output ports/roles on the trace path of those steps. On the right-hand-side of the figure we give a screenshot of the actual tags the scientist has used when **publishing** this workflow in the myExperiment workflow repository. The reader will notice that, the three steps Figure 3 correspond to those highlighted by the free-text tags "hec, hessi, and hfc".

3.1 Proposed Approach

Figure 4 illustrates the overall approach that we propose for distilling workflow provenance. The initial step is design time annotation of activities within workflow descriptions. Annotations designate the data-oriented function of the activity, such as retrieval, analysis, or data organization. We call this categorizations, motifs, which we describe in Section 3.2. Motif annotations are an important part of our approach as they enable both workflow summarization and run time origin-annotation generation. During motif annotations we also mark the significant input artifacts, such as query

[4]http://www.geneontology.org/GO.evidence.shtml

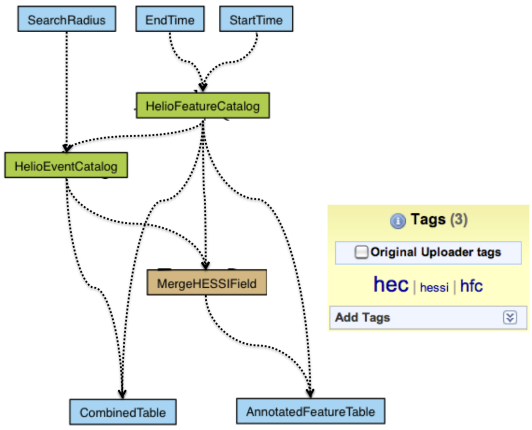

Figure 3: Summarised Form of the Heliophysics Pipeline of Figure 1.

parameters of retrieval tasks, or critical configuration parameter of an analysis task. Annotation can be a manual or a semi-automated task with the use of a classifier that mines existing workflows to suggest motif annotations.

The next step is the execution of this annotated workflow. During execution the workflow engine will specifically perform the following:

- **Acquiring Origin-Annotations on Data** : Certain activities within workflows are **opaque** processing steps that mint new data from the given inputs. In our proposal, these steps not only mint data, but also mint **origin annotations** on data. Examples of these activities are Data Retrieval and Data Analysis. Origin annotations partly tackle the black-box challenge. In our approach we do not aim for making all data processing steps transparent, instead, for those activities that mint data, we want to make qualified links between result data artifacts and the parameter/configurations that have helped *originate* (e.g. the service endpoint of the HEC retrieval step) or *scope* (e.g. query parameters of the HEC step) the result dataset.

- **Propagating Origin-Annotations**: The workflow environment that we will use, namely the Taverna workbench [22] will be extended with table-aware components, that provide a well-defined set of data organization functions such as Column Projection, Filtering and Joining. These components, we call data relaying steps, will allow us to propagate origin-annotations. The relaying is done either completely **transparently** by using table-aware components, such as a filtering step with a well-defined filter expression or **semi-transparently** such as human-editing based cleaning of data records, where the workflow environment can associate an output record with an input record but cannot tell which attributes of the record have been edited/touched during curation. In such a case record level annotations could be propagated yet the attribute level ones cannot. Assuming a structure to data and catering for data-relaying operations allow us to trace input to output value-copying relations at a fine-grain and consequently the ability to propagate annotations.

Finally, Motif annotations over the workflow description could be used to generate workflow summarizations. Summarizations help tackle the obfuscation problem by eliminating secondary steps and retaining significant steps in the workflow. The significance

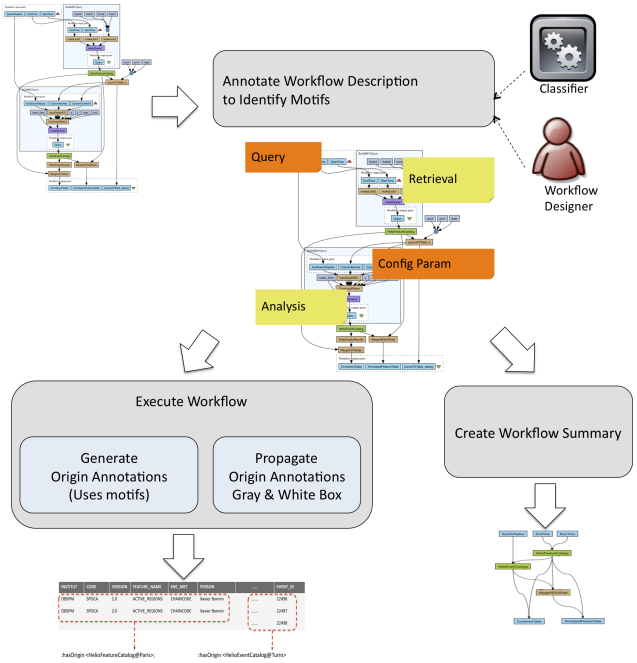

Figure 4: Overall Approach to Distilling Workflow Provenance.

markers on the input/output parameters also inform the summarization process.

3.2 Motifs in Scientific Workflows

Central to the distillation of provenance is the notion of **motifs**, which we outlined in previous work [14] based on an empirical analysis[5] of 111 scientific workflows. Motifs characterize the data-oriented nature of activities undertaken by workflows. The workflow examples presented earlier in Figure 1 demonstrates the different kinds of data motifs. The motifs are captured within a light-weight ontology accessible from[6]. In what follows, we briefly describe the data-oriented motifs:

Data Retrieval: Workflows or their certain sub-steps are used to bring-about data into the data-intensive pipeline. Data from both local and remote resources can be retrieved in various means such as queries to remote/local databases, or web service invocations such as the HEC and HFC steps in our workflow.

Data Preparation: Data access or analysis steps that are handled by external services or tools typically require well formed query strings or structured requests as input. Consequently a large number of tasks in workflows are dedicated to the generation of these queries through augmentation of multiple parameters. The BuildHFCQuery and BuildHECQuery sub-workflows are dedicated to this function in our example. The reverse operation occurs for output processing. Outputs of data access or analysis steps could be subject to data extraction or splitting to allow the conversion of data from the service specific format to the workflows internal data carrying structures (e.g. collections).

Data Organization: The data items brought into a pipeline may not be subject to analysis in their entirety. Data collections could further be filtered or could be subject to extraction of various subsets. In addition to filtering, certain tasks are dedicated to merging

[5]The analysis dataset can be found at http://www.myexperiment.org/files/789.html
[6]https://github.com/wf4ever/ro/blob/master/motifs.owl

data sets created by different branches of workflows. Examples of both filter ("FilterEmptyResults") and merge (MergeHESSIField) are present in our helio workflow.

Data Analysis/Visualization: Results of analysis processes are typically fresh piece of information that is derived from the input. In our example workflow the calculation of number of event occurances per active region is an example of such information generation.

Data Cleaning/Curation: A category not illustrated in our example. Cleaning and curation operations are typically undertaken by sophisticated tooling/services (e.g. Google Refine), or by human interactions. A cleaning/curation step essentially preserves and enriches the content of data (e.g., by a user's annotation of a result with additional information, detecting and removing inconsistencies on the data, etc.).

Data Moving: Though not exemplified in our workflow, a very common activity occurring in workflows is the movement of data in and out of the workflow environment. this is achieved through resolution of typically temporary references (e.g. downloading results from a URL, reading a file) or creation of references (e.g. uploading results to a URl or cresting a file.)

Our analysis has shown that obfuscation of workflows is a prevalent problem. More than half of all activities within workflows is related to either data-preparation or data organization. A majority of these belong to the preparation category, which are resource adapter "shim" steps that are eliminate-able during provenance distillation. Moreover in certain domains a up to one fifth of all activities are data movers which resolve or mint temporary data references not meaningful outside the workflow execution environment. This study has also informed us on the kinds of data organization constructs (e.g. filtering, joining) that should exist as part of table-aware component support in the Taverna.

4. DISCUSSION AND CHALLENGES

The re-use and re-mix of scientific datasets retrieved from community repositories into new aggregates, the application of computational analysis upon those aggregates and the publishing of the results, have become a commonplace activity in the new age of science. In this paper we provided observations on the suitability of workflow execution provenance to assist data publishing. Current workflow provenance is obfuscated and un-informative when it comes to 1) inquiring origins of data records and 2) succinctly reporting the activities within a workflow based experiment. We also provided our approach for generating provenance distillations to enable these tasks. Our provenance distillation approach raises several research issues.

Central to our approach are the motif annotations that specify the data oriented nature of workflow activities. In order to semi-automate the annotation process, one possible area of investigation is the application of data mining and machine learning techniques over workflow descriptions and workflow execution provenance.

Generation of workflow summarizations using motif annotations is another thread of work, which we anticipate, will be achieved through a set of graph-re-writing rules defined over motif combinations. As well as the above issues, we will be looking at existing graph reduction and summarization techniques, and adapt them to our problem. Our summarization approach will be applied to the local scope of a workflow (a group of workflows), in this regard our proposal is not an attempt to summarize provenance logs at large scale rather it can be seen as a preemptive noise elimination step that would allow easier aggregation of such summarizations at large scale.

We assumed the existence of a fine-grained workflow provenance model. To cater for this requirement, the work in annotation propagation in relational databases [5] has potential for applicability to our context. The empirical analysis done in the first step, however, has shown that Data Relaying activities are not always fully transparent and well-behaved, as in the case of relational query operators. Our plan in the short term is to establish a baseline provenance model and framework that caters for data (and annotation) minting opaque activities together with fully transparent data relaying activities and to demonstrate the feasibility of propagating within workflows comprised of such activities. The next step will be to incorporate the semi-transparent data relaying activities into the provenance model. We will investigate to what extent we can propagate annotations through semi- transparent activities.

Our motivation in distilling provenance comes from supporting the curator's task of experiment reporting and data publishing. Consequently, we intend to assess the effectiveness of this form of distilled provenance in the context of knowledge discovery in general, and data publishing and citation in particular. Our plan is to perform the evaluation of effectiveness of workflow summarizations and propagated annotations with users from the Biodiversity[7] and Astronomy[8] communities.

Acknowledgment

This work was supported by the myGrid platform Grant.

5. REFERENCES

[1] Recommended practices for citation of data published through the GBIF network. (May), 2012.

[2] Y. Amsterdamer, S. B. Davidson, D. Deutch, T. Milo, J. Stoyanovich, and V. Tannen. Putting lipstick on pig: Enabling database-style workflow provenance. *PVLDB*, 5(4):346–357, 2011.

[3] M. K. Anand, S. Bowers, and B. Ludäscher. Provenance browser: Displaying and querying scientific workflow provenance graphs. In *ICDE*, pages 1201–1204, 2010.

[4] R. Bentley, J. M. Brooke, A. Csillaghy, D. Fellows, A. L. Blanc, M. Messerotti, D. Perez-Suarez, G. Pierantoni, and M. Soldati. Helio: Discovery and analysis of data in heliophysics. In *eScience*, pages 248–255. IEEE Computer Society, 2011.

[5] D. Bhagwat, L. Chiticariu, W. C. Tan, and G. Vijayvargiya. An annotation management system for relational databases. In *Proceedings of the 13th VLDB Conference*, pages 900–911. Morgan Kaufmann, 2004.

[6] O. Biton, S. Cohen-Boulakia, S. B. Davidson, and C. S. Hara. Querying and Managing Provenance through User Views in Scientific Workflows. *2008 IEEE 24th International Conference on Data Engineering*, pages 1072–1081, Apr. 2008.

[7] J. Cheney, L. Chiticariu, and W. C. Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474, 2009.

[8] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *SIGMOD Conference*, pages 1345–1350, 2008.

[9] H. V. de Sompel and C. Lagoze. All aboard: toward a machine-friendly scholarly communication system. In *The Fourth Paradigm*, pages 193–199. 2009.

[7] http://www.biovel.eu/

[8] http://amiga.iaa.es/p/1-homepage.htm

[10] E. Deelman, D. Gannon, M. S. Shields, and I. Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Comp. Syst.*, 25(5):528–540, 2009.

[11] S. C. Dey, D. Zinn, and B. Ludäscher. Propub: towards a declarative approach for publishing customized, policy-aware provenance. In *Proceedings of the 23rd international conference on Scientific and statistical database management*, SSDBM'11, pages 225–243, Berlin, Heidelberg, 2011. Springer-Verlag.

[12] B. Francine. Got Data? A Guide to Data Preservation in the Information Age. *Communications of the ACM*, 51(12):50–56, 2008.

[13] M. Gamble and C. Goble. Quality, trust, and utility of scientific data on the web: Towards a joint model. In *Proceedings of the ACM WebSci'11*, Koblenz, Germany., June 2011.

[14] D. Garijo, P. Alper, K. Belhajjame, O. Corcho, C. Goble, and Y. Gil. Common motifs in scientific workflows: An empirical analysis. In *In the proceedings of the IEEE eScience Conference*. IEEE CS, 2012.

[15] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[16] D. Hull, R. Stevens, P. Lord, C. Wroe, and C. Goble. Treating shimantic web syndrome with ontologies. In *AKT Workshop on Semantic Web Services*, 2004.

[17] R. Ikeda, J. Cho, C. Fang, S. Salihoglu, S. Torikai, and J. Widom. Provenance-based debugging and drill-down in data-oriented workflows. In *ICDE 2012*. Stanford InfoLab.

[18] P. Ingwersen and V. Chavan. Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC bioinformatics*, 12 Suppl 1(Suppl 15):S3, Dec. 2011.

[19] J. Kim, E. Deelman, Y. Gil, G. Mehta, and V. Ratnakar. Provenance trails in the wings-pegasus system. *Concurr. Comput. : Pract. Exper.*, 20(5):587–597, Apr. 2008.

[20] B. F. Lavoie. Technology Watch Report The Open Archival Information System Reference Model : Introductory Guide. (January), 2004.

[21] P. Missier, S. S. Sahoo, J. Zhao, C. A. Goble, and A. P. Sheth. *Janus*: From workflows to semantic provenance and linked open data. In *IPAW*, pages 129–141, 2010.

[22] P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. A. Goble. Taverna, reloaded. In M. Gertz and B. Ludäscher, editors, *SSDBM*, volume 6187 of *Lecture Notes in Computer Science*, pages 471–481. Springer, 2010.

[23] C. Scheidegger, D. Koop, E. Santos, H. Vo, S. Callahan, J. Freire, and C. Silva. Tackling the provenance challenge one layer at a time. *Concurrency and Computation: Practice and Experience*, 20(5):473–483, 2008.