Requirements for a Provenance Visualization Component

Markus Kunde, Henning Bergmeyer, and Andreas Schreiber

Simulation and Software Technology
German Aerospace Center
51147 Cologne, Germany
{Markus.Kunde,Henning.Bergmeyer,Andreas.Schreiber}@dlr.de
http://www.dlr.de/sc

Abstract. The need for interpretation of provenance data increases with the introduction of further provenance related IT-systems. The interpretation of data only becomes intuitively with providing good and efficient visualization possibilities. During the development of general provenance visualization techniques, provenance users are classified into groups regarding their view to provenance information. The end-user requirements are evaluated on an abstract level to have a basis for research. Different intentions of end-users regarding provenance are identified and put into relationship with standard visualization types. Examples for standard visualization types are given and a brief forecast to future achievements is made.

1 Introduction

The importance of recorded provenance data will become clear during the evaluation of possible fields of application (see [1,2,3]). It is imaginable that in the next years the usage of tools including provenance technology will become mandatory in domains where the trust of information is highly crucial. Besides the recording of provenance data the interpretation of it plays a central role regarding any assertions about the past, present or future. The work represented by this paper is made up of development and evaluation of general, abstract concepts for visualization of provenance data. This analysis depends on a general approach, which can be used as a basis concept for provenance visualization in applications. The target of these visualization concepts is to provide an overview about possible general visualization alternatives.

The paper is organized as follows. Section 2 presents the motivation behind this work. In Section 3, a general user classification is made, regarding the scope of view to provenance data. Section 4 describes the transformation of user requirements into abstract types and their allocation to general visualization types. A functional classification of abstract user questions is presented. Visualization examples represent current standard visualization possibilities. In Section 5, brief examples of other projects are presented to give a first insight of possible application areas. Section 6 describes the current state of work and gives a forecast

J. Freire, D. Koop, and L. Moreau (Eds.): IPAW 2008, LNCS 5272, pp. 241–252, 2008.

to future achievements. Finally, in Section 7, a conclusion is presented including a brief evaluation of current and expected future outcomes.

2 Motivation

At the moment, the introduction of provenance on the market of IT-systems is still continuing. As the number and quality of concepts including provenance increases, further investments in the evolution of it will be made. The idea of storing provenance data grows as the concepts become more concrete and specific. This evolution comes upon its boundaries where application domain experts want to use these concepts. Storage of provenance data is one part of the whole topic, whereas the interpretation of data to get useful information is the other one.

With respect to the interpretation of provenance data the development of a provenance visualization concept becomes difficult in the manner of having a general approach for the visualization technique on the one hand and not to loose the connection to specific requirements of a concrete application domain on the other hand. The main intention of this work is to build-up general visualization concepts and their evaluation regarding concrete requirements. The advantage for the provenance community is based on the fundamental discover and development of different visualization techniques and their evaluation regarding possible application domains.

3 User Classification

The idea of analyzing provenance information depends on several circumstances like the application area of the concrete implementation and the individual task of a user of this application. The evaluation approach of these different intentions is to identify general user roles in the manner of different views to data and information and to group them together into generic user classes. Regarding the evaluation of a possible division the identified user groups are derived from the user requirements document of the EU Grid Provenance project [3].

In the context of user groups and provenance information a division between user and system provenance data is made. The term user provenance is used for workflow related provenance information. In this case the interaction-sequence with the involved user(s), the intermediate and end results and other direct workflow related information is important. The term system provenance is related to IT-system internal components and their relationship together. The exact relationship between IT-system specific components and their message exchange is mentioned with this term. The following list represents the identified abstract user roles and gives a brief explanation of each classification:

- **General User.** The general user should only see the user provenance information that is connected to workflows. The general user is involved in the configuration of the workflow. Only provenance information directly related

to the own work-surrounding field is needed. The main intention is to rely on the outcome of a workflow and to check the authentically of these results.
- **Designer.** The designer role has main access to all user related provenance data, independent of the origin, which appears in the context of the monitored system. The designer is interested in the behavior of the workflow as well as the interaction between services or the connection with the outside world.
- **Manager.** A manager can see the owned user and system provenance data. The manager monitors the provenance usage on a whole to ensure the correctness of the individual services. This role is intended to support the interpretation steps and to ensure the quality of the provenance system.
- **Administrator/Developer.** The role of the administrator or developer is designed to capture the whole provenance data, which is available in the connected provenance stores. The purpose of this role is to build-up the provenance architecture and to ensure the correctness of the provenance system.

4 Generalized User Requirements

For the development of visualization concepts, a clear understanding of users need and users view regarding provenance information is mandatory. For evaluation of a general visualization concept, a derivation of user requirements for a special application must be made in order to have a universal assertion as a basis for these concepts. This is done by derivation of abstract types of identified user requirements. These types present the general intent of a user regarding provenance visualization. Besides the types of user requirements there is a need for a definition of an abstract layer of all provenance questions in relation to their point of interest.

4.1 Types

The derivation of types of user requirements into a more abstract view in order to display a general division of non-concrete user requirements is formed in the context of what element is the basis for visualization. Visualization is based on one element, the point of interest with additional information, with respect to the provenance data. The fundamental user requirements are extracted from [6]. The general approach for the derivation of the types was a two-way strategy. At first a bottom-up approach was used for a pre-selection of types. The pre-selection then was transformed into type-categories. Finally, a top-down approach was used to divide the user requirements into each type-category. In a further step, these types can be assigned to general visualization types, which were used as an essential for developing concrete visualization-concepts (see also 4.3).

Table 1 displays the abstract types of user requirements in which a user requirement can be arranged with a very brief denotation of each type.

Table 1. Types of user requirements

Type	Denotation
Process	In the center of the users view the process plays the central role. The approach of a workflow has to be evaluated. Involved actors as well as their connection are important. The sequence of the process steps is in the center of inspection.
Results	The intermediate or end results of interactions are in the center of users view. The outcome as well as the input has to be evaluated.
Relationship	In this case the relationship of interactions or actors is important and has to be evaluated. It is mandatory to reconstruct the evolution process of a result for reliance, in order to evaluate the results properly.
Timeline	If the time is important to observe, finding bottlenecks or trying of improvement of the workflow is one of the targets. Reconstructing the evolution of results or the behavior of actors to each other can be evaluated.
Participation	The evaluation of the correctness of the participants is important in the context of trust of the data. This type is very similar to the type Relationship, but there is another intention. The reconstructing of evolution processes is less important than the trust of all participated actors, which is mentioned with this type.
Compare	The comparison of two subjects deals with the differences between them. In the case of a comparison between one subject and a reference subject, the correctness of the subject can be proven.
Interpretation	This type represents a collection of individual questions, which cannot be classified into one of the other types. This type is represented with an individual visualization view depending on the special question of the end-user. Typical examples for these types are user requirements tend to develop new cognitions onto existing information.

4.2 Classification

As the division into types of user requirements is made to have an abstract division for assigning to basic visualization possibilities, a classification of the user requirements in the context of the user questions (listed in table 2) can be made. A classification of user requirements represents a functional division of user requirements. This division can be used to evaluate the fundamental provenance data, which is needed in order to give an answer to the user questions. Table 2 lists each classification and gives the abstract question behind it. Each user requirement related to the interpretation of the provenance should belong to this classification.

At first glance, there is interference between the classification and the division into types of the user requirements. The division into these two fields is made because of the different view of each field. The division into types is made in context of a possible visualization-panel in opposition to the classification, which context is the intention of user's question.

Table 2. Classification of user requirements

Classification	Abstract Formulation
Question of origin	What data was used in the generation of a data item?
Question of inheritance	What data items and information were generated using a given data item?
Question for participants	Which actors (users, applications, versions of tools, etc.) were employed in the generation of a data item?
Question for dependencies	Which resources from other projects/processes have been used in the generation of a data item?
Question for progress	In what stage of a processing chain is a given data item (for data items of the same type)? Has the process the data item is part of been finalized?
Question for quality	Did the process the data item is part of reach a satisfactory conclusion by some given regulations or criteria?

4.3 Visualization

At this point a rough assertion of visualization concepts is displayed (regarding process in [5]). This listing is made with the intention to have contrasting visualization domains, which are asserted to standard visualization concepts. These were fundamental for the ongoing project.

The general approach regarding the development of visualization concepts is represented by four steps. At first the user requirements are categorized into types. Secondly, existing standard visualization types are evaluated. The next step contains a matching between the user requirement types and the evaluation of standard visualizations. Regarding the results of the previous steps the concrete visualization concepts are developed.

Table 3 lists generic visualization types, allocates them with user requirements types and requirement classifications and briefly describes them. The type 'Interpretation' is missing in table 3. It is arguable if interpretation is carried out in every visualization type but primarily interpretation is completed by the user.

4.4 Visualization Examples

In this section few visualization examples are displayed (in addition to [7]). They depend on the division of visualization assertions and represent a first assertion of provenance information and their representation in standard visualization types, which will be evaluated to final visualization concepts. Each visualization example, representing only a first abstract sketch, evaluates the visualization technique in the context of one point of interest. After development of the final concepts, a complete evaluation of each visualization proposal will be made. The manipulation of information (e.g. zoom function of detail depth, filtering or sorting) is not considered in the sketches, but will be considered in the final visualization concepts. Regarding the detail level and scope of each individual visualization technique a visualization map (describes the behavior and relationship

Table 3. Visualization assertions

Visualization Type	**Description**	**Related Type**	**Related Classification**
Process diagram	The process diagram highlights the workflow with its actors, interactions and results into the center of users view.	Process Results Participation	Participants Dependencies Progress
Difference diagram	The difference diagram displays the difference between the compared objects (process, actor, interaction).	Compare	Quality
Dependency diagram	The dependency diagram displays the connection of the chosen elements (e.g. actors, interactions). It presents the behavior and the relation between income and outcome to each other.	Results Relationship Participation	Origin Inheritance Participants Dependencies
Timeline diagram	The timeline diagram displays all interactions between actors in the context of their relationship in a timeline. This diagram is similar to the process diagram, but in this diagram qualified connections are displayed.	Process Results Relationship Timeline Participation	Origin Inheritance Progress
Spreadsheet representation	The spreadsheet representation gives the most space for doing interpretation of the data. In order to have full freedom for sorting and filtering elements, this is the most flexible but also the most unclear representation strategy.	Process Results Relationship Timeline Participation Compare	Quality

of visualization concepts to each other) will be developed with respect to the scalability of the visualizations.

All visualization examples describe the three-way handshaking (or a part of it) used in information technology or related fields.

Flow Chart (Related to Process Diagram)**.** The flow chart diagram represents the visualization of the complete workflow in the context of having actors interacting to each other and related data. This diagram type is intended for representation of interactions. The key points of interest are: process sequence, combination of actors and interactions, who interact with whom?, input and outcome of an actor, data transformation. The complete three-way handshaking is displayed in the example with focus on actors and data.

Data Flow Diagram (Related to Process Diagram)**.** The data flow diagram represents the visualization of the complete workflow in the context of having actors interacting to each other and related interaction sequences. This diagram

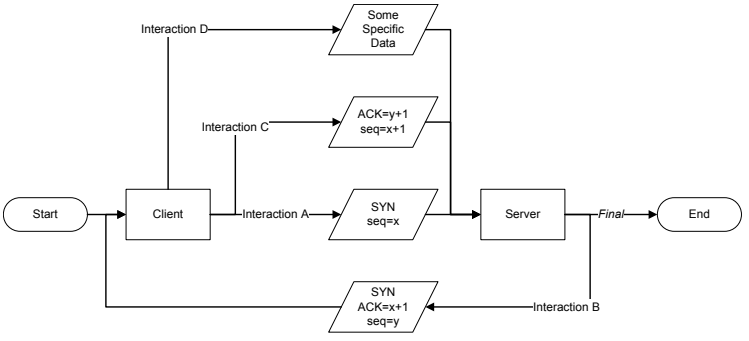

Fig. 1. Sketch of flow chart visualization type

type is intended for representation of actors and their interactions. The key points of interest are: information flow sequence, interaction-call sequence, who interact with whom?, factual process sequence. The complete three-way handshaking is displayed in the example with focus on actors.

Difference Diagram (Related to Difference Diagram)**.** The difference diagram compares a workflow or data with comparable data. Differences are highlighted. The key points of interest are: comparison of two objects (processes, data, actor states, interactions). The complete three-way handshaking is compared with a reference workflow. The difference is highlighted in the example.

System Context Diagram (Related to Dependency Diagram)**.** The system context diagram displays a central point (e.g. a workflow, interaction, actor or data) and the relation to any other part. This diagram type is intended for representation of relationships and states. The key points of interest are: effecting

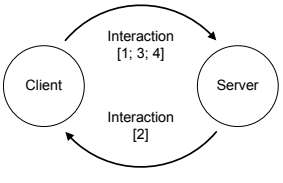

Fig. 2. Sketch of data flow diagram visualization type

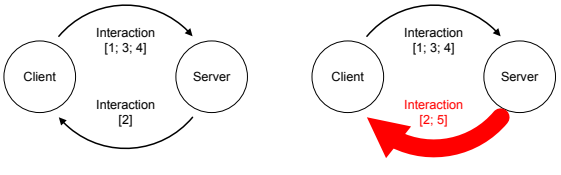

Fig. 3. Sketch of difference diagram visualization type

relationships achieved, effecting relationships published, input and outcome of an actor. The complete three-way handshaking is displayed in the example with focus on one actor and its relationships to other involved elements regarding the direction of their impact.

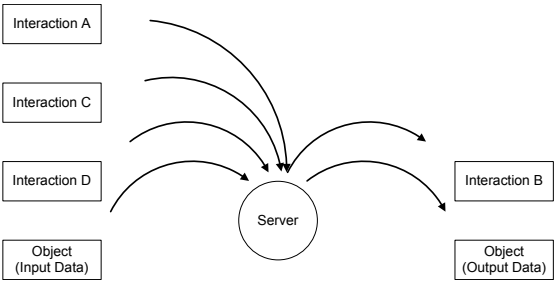

Fig. 4. Sketch of system context diagram visualization type

Brainstorm Diagram (Related to Dependency Diagram)**.** The brainstorm diagram represents any related content regarding a central point. It displays all elements which have an effect to the central point or where the point has an effect to. This diagram type is intended for representation of relationships. The key points of interest are: relationship of input and outcome (data, interactions, actors). The complete three-way handshaking is displayed in the example with focus on one interaction and its relationships to other involved elements.

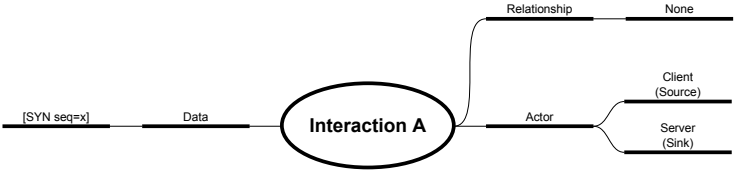

Fig. 5. Sketch of brainstorm diagram visualization type

Fishbone Diagram (Related to Dependency Diagram)**.** The fishbone diagram, also known as cause-and-effect diagram, displays any related causes to a point. This diagram type is intended for representation of relationships. The key points of interest are: relationship of input and outcome (data, interactions, actors). The complete three-way handshaking is displayed in the example with focus on the impact of elements.

State Chart Diagram (Related to Dependency and Timeline Diagram)**.** The state chart diagram displays all states of an actor during the life-cycle of a

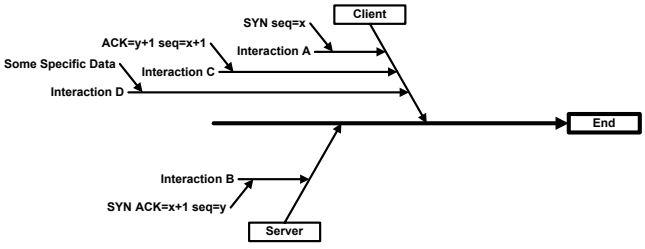

Fig. 6. Sketch of fishbone diagram visualization type

workflow. The key points of interest are: actor states, transforming interactions, transformed data, time-context. The complete three-way handshaking is displayed in the example with focus on state changes and their 'appearance-chain'.

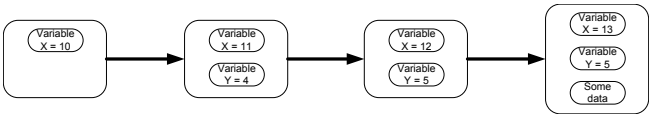

Fig. 7. Sketch of state chart diagram visualization type

Sequence Diagram (Related to Timeline Diagram)**.** The sequence diagram represents the sequence of interactions of related actors in the context of a timeline. This timeline can be qualified or unqualified. This diagram type is intended for representation of interactions and states in a time-context. The key points of interest are: time-context of process, involved actors, executed interactions, input and outcome data. The complete three-way handshaking is displayed in the example with focus on actors and interactions.

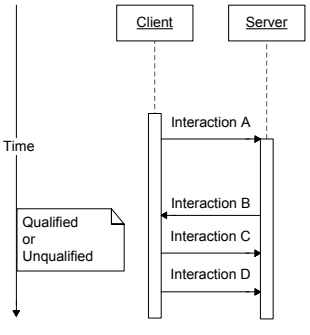

Fig. 8. Sketch of sequence diagram visualization type

ProcessStep	Type	Interaction-Name	MessageSource	MessageSink	Content
1	Interaction	Interaction A	Client	Server	SYN seq=x
2	Interaction	Interaction B	Server	Client	SYN ACK=x+1 seq=y
3	Interaction	Interaction C	Client	Server	ACK=y+1 seq=x+1
4	Interaction	Interaction D	Client	Server	some specific data
5	Relationship	Interaction B	Interaction A		causedBy
6	Relationship	Interaction C	Interaction B		causedBy
7	Relationship	Interaction D	Interaction B		causedBy
8	Actorstate	Interaction A	Client		X=10
9	Actorstate	Interaction B	Server		X=11; Y=4
10	Actorstate	Interaction C	Client		X=12; Y=5
11	Actorstate	Interaction D	Client		X=13; Y=5; [some specific data]

Fig. 9. Sketch of spreadsheet visualization type

Spreadsheet (Related to Spreadsheet Representation)**.** The spreadsheet representation displays a scheduler collection of provenance datasets with the possibility of filtering or sorting of the results. This representation type is intended for detailed information research. The key points of interest are: Displaying all relevant information (interactions, relationships, actor states, time-context). The complete three-way handshaking is displayed in the example.

5 Examples from Projects

This section covers a selection of possible applications regarding provenance visualization. These examples already use the provenance system or are a proper candidate for employment. As it is obvious all applications use the provenance technology in a different way. In some cases provenance is used to understand the behavior of the IT-system (e.g. TENT) while other systems' usage is (partly) based on provenance (e.g. ENCHR, VisTrails).

C3-Grid. The main goal of the C3-Grid project is to do research about the earth system for understanding the behavior and dynamic of the whole and each subsystem [4]. The verification of the model and the data of this simulation is one possible application point for a visualization concept based on the result of this project.

TENT. TENT [8] is a software integration and workflow management system that simplifies work by building up simulation process chains in distributed environments. The visualization concepts provide a graphical way for evaluation of the workflows regarding increased quality and trust of the outcome [9].

ENCHR. The 'Electronic Healthcare Record System' (ENCHR) is a solution for an unbound healthcare situation [3]. The traceability and trust of each result is mandatory. With adequate concepts for a visual interpretation of this evolution a fast and correct consequence can be covered.

OTM. One further example for a possible application is the 'Organ Transplant Management' (OTM), already mentioned as a prime example in the provenance project [3]. A sophisticated visualization concept supports the tasks regarding the diversified group of possible provenance users.

VisTrails. The software 'VisTrails' is a good example to explain the need of good visualization concepts for interpretation of provenance data [10]. One intention of the software is to support an expert in data exploration, the systematic tracking of workflow evolution and to comprehend the steps made. The software shows the advantages of a good visualization concept supporting data interpretation.

6 Current and Future Work

Currently, the analysis of user requirements is done and its division into abstract types and its classification. Possible end-users are identified and grouped into different user roles. First assertions about standard visualization types, matching to the abstract intends of the users, are made and evaluated [5]. During next project phases the existing standard visualization types are being evaluated in more detail to enhance them into concrete visualization concepts [5]. New Visualization approaches are being developed regarding modern visualization techniques, such as tree maps, magic lens, network visualization and others. These visualization concepts will be evaluated regarding users' requirements.

7 Conclusions

The need for interpretation and visualization of provenance data increases step-by-step by the ongoing development of provenance technology and its introduction in real IT-systems [11]. The increasing number of application areas surrounds the usage and analysis of provenance data from application domain experts. Regarding this evolution the analysis of provenance data should become more easy and intuitive; considering the background of each application domain and the intention of the operating end-user. In this paper a first insight into the visualization of provenance data is given. A classification of user and requirements is made and a first assertion about possible visualization types is presented. With respect to other research projects [2,3], which evaluates the need and a concrete application for querying and exploring of provenance data, this paper describes an approach for visualization of these steps, taking a further step in the direction to end-users.

Acknowledgments. This work has been supported by the German Federal Ministry for Research and Technology (BMBF) under Grant 01IG07006A.

References

1. Groth, P., Jiang, S., Miles, S., Munroe, S., Tan, V., Tsasakou, S., Moreau, L.: An architecture for provenance systems. Technical report, Provenance Consortium (2006)
2. The Pasoa Website, `http://www.pasoa.org`
3. The EU Grid Provenance Project Website, `http://www.gridprovenance.org`

4. The C3-Grid Website, http://www.c3grid.de
5. Fry, B.: Visualizing Data, 1st edn. O'Reilly, Sebastopol (2007)
6. WorkPackage2: Grid provenance user requirements document. Technical report, Provenance Consortium (2005)
7. Deora, V., Contes, A., Rana, O.: Tool for Navigating Provenance Information. In: Provenance Challenge Workshop, Cardiff University (2006)
8. Schreiber, A.: The integrated simulation environment TENT. Concurrency and Computation: Practice and Experience (13-15), 1553–1568 (2002)
9. Kloss, G.K., Schreiber, A.: Provenance implementation in a scientific simulation environment. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 37–45. Springer, Heidelberg (2006)
10. Freire, J., Silva, C.T., Callahan, S.P., Santos, E., Scheidegger, C.E., Vo, H.T. (eds.): Managing rapidly-evolving scientific workflows, University of Utah (2006)
11. Miles, S. (ed.): Electronically Querying for the Provenance of Entites, School of Electronics and Computer Science, University of Southampton (2006)