Data Lineage, Workflows, and Science Products

Albrecht Schmidt Gerhard Schwehm

Interdepartmental Workshop, ESTEC, 28/29 July 2008

eScience Projects

- Identify Problem

eScience Projects

- Identify Problem
- Hypothesis

eScience Projects

- Identify Problem
- Hypothesis
- Test Hypothesis

eScience Projects

- Identify Problem
- Hypothesis
- Test Hypothesis
- Analyse Results

eScience Projects

- Identify Problem
- Hypothesis
- Test Hypothesis
- Analyse Results
- Draw Conclusions

eScience Projects

- Identify Problem
- Hypothesis
- Test Hypothesis
- Analyse Results
- Draw Conclusions
- Identify Next Problem ...

Role of Data in eScience Projects

- Explorative experimentation

Role of Data in eScience Projects

- Explorative experimentation
- Interactive Environments

Role of Data in eScience Projects

- Explorative experimentation
- Interactive Environments
- Traditional auditing methods fail to be implemented

Role of Data in eScience Projects

- Explorative experimentation
- Interactive Environments
- Traditional auditing methods fail to be implemented
  - Sometimes, they *cannot* be implemented.

Role of Data in eScience Projects

- Explorative experimentation
- Interactive Environments
- Traditional auditing methods fail to be implemented
  - Sometimes, they *cannot* be implemented.
- Consequently, data quality is difficult to establish

Role of Data in eScience Projects

- Explorative experimentation
- Interactive Environments
- Traditional auditing methods fail to be implemented
  - Sometimes, they *cannot* be implemented.
- Consequently, data quality is difficult to establish
- Alleged trends towards data-driven science – data deluge (?)

Technical Challenges

Technical Challenges

- Seemingly not many – but integration is the main issue.

Technical Challenges

- Seemingly not many – but integration is the main issue.
  - Audit trails
  - Pattern mining
  - Temporal transactional logic, temporal databases
  - Configuration Control
  - ...

Technical Challenges

- Seemingly not many – but integration is the main issue.
  - Audit trails
  - Pattern mining
  - Temporal transactional logic, temporal databases
  - Configuration Control
  - ...
- However:

Technical Challenges

- Seemingly not many – but integration is the main issue.
  - Audit trails
  - Pattern mining
  - Temporal transactional logic, temporal databases
  - Configuration Control
  - ...
- However:
  - Often great part of the burden on the user
    - Need to learn new concepts, APIs, ...
  - Impact on workflow

Impedance Mismatches

- Data useful at different stages for different purposes
  - Can often do useful science with non-committed data
- Audit requirements interfere with:
  - Research processes
  - Skill sets
  - Additional complexity often not immediately visible
- Adds layers of complexity
  - Research processes
  - Skill sets
  - Additional complexity often not immediately visible

Knowledge Managment on the Web 2.0 – Relax Processes

- Web 2.0 phenomena like Wikis, Tagging, . . .
  - Expand on analysis of hyperlinks.
  - Leverage even more user interaction.

Knowledge Managment on the Web 2.0 – Relax Processes

- Web 2.0 phenomena like Wikis, Tagging, . . .
  - Expand on analysis of hyperlinks.
  - Leverage even more user interaction.
- Common Trait: Relax Processes – lower entry thresholds
  - Infer knowledge from user behaviour
  - Encourage auditing of inferred knowledge

Knowledge Managment on the Web 2.0 – Relax Processes

- Web 2.0 phenomena like Wikis, Tagging, . . .
  - Expand on analysis of hyperlinks.
  - Leverage even more user interaction.
- Common Trait: Relax Processes – lower entry thresholds
  - Infer knowledge from user behaviour
  - Encourage auditing of inferred knowledge
- Take into account observed user behaviour
  - Infer knowledge from user behaviour
  - Encourage auditing of inferred knowledge

Re-phrase Some of the Principles

- User behaviour: logs and audit trails
  - Use history, trace files, monitors, ...
  - Exploit wisdom of crowds?

Re-phrase Some of the Principles

- User behaviour: logs and audit trails
  - Use history, trace files, monitors, ...
  - Exploit wisdom of crowds?
- Automatically create documentation

Re-phrase Some of the Principles

- User behaviour: logs and audit trails
  - Use history, trace files, monitors, ...
  - Exploit wisdom of crowds?
- Automatically create documentation
- However:
  - User has to review
  - User has to authorise

Re-phrase Some of the Principles

- User behaviour: logs and audit trails
  - Use history, trace files, monitors, ...
  - Exploit wisdom of crowds?
- Automatically create documentation
- However:
  - User has to review
  - User has to authorise
- Additionally: Employ low-threshold tools
  - Encourage use of wikis, blogging platforms, ...

Future Work

- Try out these ideas
  - Make sure the community's approach is respected.
  - Take into account the size of the community.
- Find out about usability and user acceptance
  - Avoid patronising the user.
    - This includes the requirement to learn complex techniques and the avoidance of radical adaption of processes.
  - Rely on standards.