

Data Provenance-Tools

KSWS/Projekt/NEidl WiSe 2020/21

Betreuer:

- Tanja Auge

Beschreibung:

Provenance Management beschreibt die Herkunft eines konkreten (meist digitalen) Ergebnisses zurück zu seinen physikalischen Wurzeln. Im Bereich Provenance unterscheiden wir dabei zwischen verschiedenen Provenance-Typen: Die wichtigsten sind das Aufspüren eines konkreten Ergebniswertes (Data Provenance), das Sammeln von Metadaten über den Produktionsprozess (Metadata Provenance) oder den Prozess selbst (Workflow Provenance). Ziel des Projekts ist die Klassifizierung und Analyse theoretisch oder praktisch besonders relevanter Tools im Bereich Data Provenance, welche in den letzten 30 Jahren entwickelt wurden. Hierzu zählen unter anderem ORCHESTRA, Perm und Trio. Einige ausgewählte Tools wurden im letzten Semester auf ihre verschiedenen spezifischen Anwendungsfälle analysiert, ihre Funktionalität untersucht und die zugrundeliegende Theorie an einem konstanten Studenten-Beispiel erklärt. Im Folgeprojekt soll diese Auswahl noch ein wenig erweitert und alle bisherigen Analysen in einem gemeinsamen wissenschaftlichen Artikel beschrieben werden. Ziel ist somit die Erstellung einer wissenschaftlichen Veröffentlichung.

Arbeitsschritte:

- Einarbeiten in das Thema Data Provenance sowie aktuelle Vorarbeiten
- Literaturrecherche:
 - Was ist Data Provenance?
 - Welche Tools existieren bereits im Bereich Data Provenance?
 - Welche Tools sind besonders wichtig/bekannt?
- Erweiterung der Tooltests:
 - Analyse neuer Tools wie ProvC&B
 - Aufbereitung der Tools als virtuelle Maschinen
- Verfassen der schriftlichen Ausarbeitung:
 - Beschreibung der Grundlagen
 - Beschreibung der Tools
 - Erstellung einer wissenschaftlichen Veröffentlichung

Voraussetzungen:

- Datenbanken I
- Vorwissen im Bereich Data Provenance

Literatur (siehe Stud.IP):

- Provenance Systems Roadmap (Pérez, Rubio, Sáenz-Adán)
- A systematic review of provenance systems (Pérez, Rubio, Sáenz-Adán)
- Provenance in Databases: Why, How, and Where (Cheney, Chiticariu, Tan)
- A survey on provenance (Herschel, Diestelkämper, Lahmer)
- Navigator — Ursprung von Daten (Sila)
- Abschlussbericht der KSWS-Gruppe SS2020 (Flach, Herrmann, Kaseler, Röhrs, Strelnikov)

Prüfung:

- Präsentation des aktuellen Zwischenstandes (1x pro Monat, 10 min)
- Abschlusspräsentation:
 - Vortrag (20 min) sowie
 - Schriftliche Ausarbeitung (10 Seiten)

Meilensteine:

- 10.11.2020: Auftakttreffen
- 01.12.2020: Vorstellung der Literaturanalyse sowie Auswahl der Tools (Meilenstein I)
- 05.01.2021: Definition der Grundstruktur der Artikels sowie Analyse der neuen Tools (Meilenstein II)
- 02.02.2021: Schriftliche Ausarbeitung der Grundlagen und Aufbereitung der virtuellen Maschinen (Meilenstein III)
- 02.03.2021: Schriftliche Ausarbeitung der Tool-Berichte (Meilenstein IV)
- xx.03.2021: Abschlusspräsentation (Meilenstein V)