LTAT.05.008: Software Analytics

Case study research

Dietmar Pfahl
email: dietmar.pfahl@ut.ee

Spring 2018
Quiz 2: Surveys

(10 min -- 2 marks)
Data Collection & Research Methods in Data Science

• Survey
  – Questionnaire-based (primary study)
  – Literature-based (secondary / tertiary study)

• Case Study
  – Exploratory
  – Descriptive
  – Confirmatory/Explanatory
  – Improving

• Experiment
  – Controlled Experiment
  – Quasi-Experiment

• Many other…
  – Action Research
  – Ethnography
  – Longitudinal Studies
  – Design Science

Data Mining
Case Study
-- Definition (Social Science)

“Case study is an empirical enquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident.”

Source:
Case Study
-- Definition (Information Systems)

“A case study examines a phenomenon in its natural setting, employing multiple methods of data collection to gather information from one or a few entities (people, groups, or organization). The boundaries of the phenomenon are not clearly evident at the outset of the research and no experimental control or manipulation is used.”

Source:
Case Study
-- Definition (Software Engineering)

“Case study in software engineering is an empirical enquiry that draws on multiple sources of evidence to investigate one instance (or a small number of instances) of a contemporary software engineering phenomenon within its real-life context, especially when the boundary between phenomenon and context cannot be clearly specified.”

Source:
Case Study – Other terms used …

• Field Study / Observational Study
• Archival Study / History Study
  – if large sets of data collected in the past are analysed (data mining)
• Action Research
  – if researcher is involved in change/improvement activities
• Ethnography
  – if main focus is on cultural aspects
Case Study & Research Strategy

• **Exploratory** – finding out what is happening, seeking new insights, and generating ideas and hypotheses for new research

• **Descriptive** – portraying the current status of a situation or phenomenon.

• **Explanatory (Confirmatory)** – seeking an explanation for a situation or a problem, mostly but not necessarily, in the form of a causal relationship (or: seeking a confirmation of a theory).

• **Improving** – trying to improve a certain aspect of the studied phenomenon.
Case Study & Research Strategy

• **Exploratory** – finding out what is happening, seeking new insights, and generating ideas and hypotheses for new research

• **Descriptive** – portraying the current status of a situation or phenomenon.

• **Explanatory (Confirmatory)** – seeking an explanation for a situation or a problem, mostly but not necessarily, in the form of a causal relationship (or: seeking a confirmation of a theory).

• **Improving** – trying to improve a certain aspect of the studied phenomenon.
Case Study & Triangulation

- **Triangulation** is important to increase the precision and strengthen the validity of empirical research

- Types of Triangulation:
  - **Data (Source)** Triangulation – using more than one data source or collecting the same data at different occasions
  - **Observer** Triangulation – using more than one observer in the study
  - **Methodological** Triangulation – combining different types of data collection methods, for example, qualitative and quantitative methods
  - **Theory** Triangulation – using alternative theories or viewpoints
Case Study – Research Process Steps

1. **Case study design** – objectives are defined and the case study is planned.

2. **Preparation for data collection** – procedures and protocols for data collection are defined.

3. **Collecting evidence** – data collection procedures are executed on the studied case.

4. **Analysis of collected data** – data analysis procedures are applied to the data.

5. **Reporting** – the study and its conclusions are packaged in feasible formats for reporting.
## Case Study -- Design

<table>
<thead>
<tr>
<th>Element</th>
<th>Example Questions Describing the Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rationale</td>
<td>Why is the study being done?</td>
</tr>
<tr>
<td>Purpose</td>
<td>What is expected to be achieved with the study?</td>
</tr>
<tr>
<td>The case</td>
<td>Overall, what is being studied?</td>
</tr>
<tr>
<td>Units of analysis</td>
<td>In more detail, what is being studied?</td>
</tr>
<tr>
<td>Theory</td>
<td>What is the theoretical frame of reference?</td>
</tr>
<tr>
<td>Research questions</td>
<td>What knowledge will be sought or expected to be discovered?</td>
</tr>
<tr>
<td>Propositions</td>
<td>What particular (causal) relationships are to be investigated?</td>
</tr>
<tr>
<td>Define concepts and measures</td>
<td>How are entities and attributes being defined and measured?</td>
</tr>
<tr>
<td>Methods of data collection</td>
<td>How will data be collected?</td>
</tr>
<tr>
<td>Methods of data analysis</td>
<td>How will data be analyzed?</td>
</tr>
<tr>
<td>Case selection strategy</td>
<td>How will cases (and units of analyses) be identified and selected?</td>
</tr>
<tr>
<td>Data selection strategy</td>
<td>How will data be identified and selected?</td>
</tr>
<tr>
<td>Replication strategy</td>
<td>For example, who will be interviewed? What electronic data sources are available for use in the study?</td>
</tr>
<tr>
<td>Quality assurance, validity</td>
<td>Is the study intended to literally replicate a previous study, or theoretically replicate a previous study; or is there no intention to replicate?</td>
</tr>
<tr>
<td>and reliability</td>
<td>How will the data collected be checked for quality? How will the analysis be checked for quality?</td>
</tr>
</tbody>
</table>

Source:
## Case Study -- Design

### Table: Example Questions Describing the Element

<table>
<thead>
<tr>
<th>Element</th>
<th>Example Questions Describing the Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rationale</td>
<td>Why is the study being done?</td>
</tr>
<tr>
<td>Purpose</td>
<td>What is expected to be achieved with the study?</td>
</tr>
<tr>
<td>The case</td>
<td>Overall, what is being studied?</td>
</tr>
<tr>
<td>Units of analysis</td>
<td>In more detail, what is being studied?</td>
</tr>
</tbody>
</table>

### Diagram: Single-case study vs. Multiple-case study

- **Single-case study**
  - **Context**
    - Case = unit of analysis

- **Multiple-case study**
  - **Context 1**
    - Case 1 = unit of analysis 1
  - **Context 2**
    - Case 2 = unit of analysis 2

---

**Source:**
Example:
Imagine an XP case study where two projects are studied in two different companies in two different application domains, both using Agile practices.

What is context? -- What is unit?
Option 1:
The projects may be considered two units of analysis in an embedded case study if the context is software companies in general and the research goal is to study Agile practices.

Source:
Option 2: If the context is considered to be the specific company or application domain, the two projects would be better studied as two separate holistic cases.

Source:
Research Questions

“Research questions are statements about the knowledge that is being sought, or is expected to be discovered, during the case study. The discovery or attainment of this knowledge demonstrates that the case study has achieved its intended objectives.”

Source:
Research Questions, Propositions, Hypotheses – Example:

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Proposition with Accompanying Hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1: How do motivational factors encourage longevity in clinical expert systems?</td>
<td>RP1.1: IF there are motivational factors THEN users will continue to use the clinical expert system.</td>
</tr>
<tr>
<td></td>
<td>H1.1: Motivating factors are highly correlated with the continued use of a clinical laboratory expert system.</td>
</tr>
<tr>
<td>RQ2: How do long-lasting clinical laboratory expert systems support maintenance in order to adapt to changes over time?</td>
<td>RP2.1: Long-lasting clinical expert systems are easy to maintain.</td>
</tr>
</tbody>
</table>

Source:
The Wallace Model

Theories

Hypotheses (Research Questions)

Research Design

Data Analysis, Parameter Estimation

Empirical Generalizations (Laws)

Observations

Research Methods

Logical Inference (deduction)

Theory – Characterisation

A scientific theory
– identifies and defines a set of phenomena, and makes assumptions about those phenomena and the relationships between them.
– precisely defines the theoretical terms, so that a community of scientists can observe and measure them.
– explains why certain relationships occur.

Positivists expect their theories to have strong predictive power, and so look for generalised models of cause-and-effect as the basis for theories.

Constructivists expect theories to strengthen their understanding of complex situations, and so tend to make more use of categorisations and analogies.
Theory – Exercise

• Imagine a theory related to using UML diagrams

• What could such a theory be?

“I think you should be more explicit here in step two.”
Theory – Exercise

• Imagine a theory related to using UML diagrams

• What could such a theory be?

*Hint: For example, think of "a picture tells more than 1000 words".*
Theory – Example

• A researcher might develop a theory around using UML diagrams as standardised forms of external memory.

• According to this theory, UML diagrams are used to
  – summarize the results of meetings and discussions
  – remind participants of a shared understanding that they have already developed
  – support all kinds of development activities

“I think you should be more explicit here in step two.”
Theory – Example

• A researcher might develop a theory around using UML diagrams as standardised forms of external memory.

• According to this theory, UML diagrams are used to
  – summarize the results of meetings and discussions
  – remind participants of a shared understanding that they have already developed
  – support all kinds of development activities

• The theory
  – should precisely define the meanings of terms such as “diagram”, “participant”, “discussion”, in order to identify them in any studies performed. (Constructs)
  – should be able to predict qualities of the diagrams that a team produces based on certain factors, and/or predict the quality of the software produced based on the use of UML diagrams. (Propositions)
  – should explain why developers chose to use UML diagrams in some circumstances but not in others, and why they include certain things in their diagrams and exclude others. (Explanations)
Theory – Example (cont’d)

Example of a theory description (concept):

<table>
<thead>
<tr>
<th>Constructs</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 – UML-based development method</td>
</tr>
<tr>
<td>C2 – Costs (total number of person hours in the project)</td>
</tr>
<tr>
<td>C3 – Communication (ease of discussing solutions within development teams and in reviews)</td>
</tr>
<tr>
<td>C4 – Design (modularity, structural properties of the code)</td>
</tr>
<tr>
<td>C5 – Documentation (the documentation of the system for the purpose of passing reviews as well as for expected future maintainability)</td>
</tr>
<tr>
<td>C6 – Testability = the ability to develop effective tests efficiently</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Propositions</th>
</tr>
</thead>
<tbody>
<tr>
<td>P5 – The use of a UML-based development method positively affects testability</td>
</tr>
<tr>
<td>P6 – The positive effects of UML-based development are reduced if training is not sufficient and adapted</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5 – Test cases based on UML models are easier to develop / Can be developed earlier / Are more complete / ...</td>
</tr>
</tbody>
</table>

| Source: |
| Guide to Advanced Empirical Software Engineering |
| Shell, Forrest; Singer, Janice; Staberg, Dag I. K. (Eds.) |
| 2008, XII, 385 p. 37 illus., Hardcover |
| ISBN: 978-3-4862-0543-8 |
How to get a Theory?

Theories can come “from anywhere” and be based on logical inference (deduction) or singular observations (induction).

**Grounded Theory** is a technique for developing theory iteratively from qualitative data (e.g., from interviews).

– Initial analysis of the data begins without any preconceived categories.

– As interesting patterns emerge, the researcher repeatedly compares these with existing data, and collects more data to support or emerge the supporting theory.
Grounded Theory

Code 1 ➔ Category 1 ➔ Pattern 1 ➔ Grounded Theory

Code 2 ➔ Category 2 ➔ Pattern 2 ➔ Grounded Theory

Code 3 ➔ Category 3 ➔ Grounded Theory

Code 4 ➔ Category 4 ➔ Grounded Theory

Code 5 ➔ Category 5 ➔ Grounded Theory

Code 6

Code 7

Code 8

Code 9
Open Coding

Body Adaptation, Construction, Comparative Analysis, Development, Drug, Email, Experience, Negative Response, Parents, Peer Acceptance, Programming, School, Study, Social Act, Telephone Conversation, Text Message, Tempting Fate, University, Voice Mail,

Communication
Open Coding

From the above set of codes, we can group the concepts: ‘Email’, ‘Telephone Conversation’, ‘Text message’ & ‘Voice Mail’ into a category and name it ‘Communication’.
Constructs and Measures

• The concepts being used in the research questions, propositions and hypotheses need to be defined. ➔ Constructs

• Where appropriate, measures for those concepts also need to be defined. ➔ (direct) Measures

• Some concepts may not be measurable directly and it will be necessary to identify and define surrogate measures. ➔ (indirect) Measures

• Often measures are derived from other measures (e.g., productivity) ➔ (derived) Measures
Measurable Entities in a SW Process

An entity can represent any of the following:

- **Process/Activity**: any activity (or set of activities) related to software development and/or maintenance (e.g., requirements analysis, design, testing) – these can be defined at different levels of granularity

- **Product/Artifact**: any artifact produced or changed during software development and/or maintenance (e.g., source code, software design documents)

- **Resources**: people, time, money, hardware or software needed to perform the processes
Process Measure – Sprint Productivity

Name: ’Average Sprint Productivity’ (~ ’Process Efficiency’)

Name: Avg. Sprint Prod.  ???????? ...

Entity: Software
Attribute: Sprint
Attribute: Size
Attribute: Effort
Unit: 
Range: [0, \infty)

Productivity = \frac{Output}{Input}
Process Measure – Sprint Productivity

Name: ’Average Sprint Productivity’ (~ ’Process Efficiency’)

Name: Avg. Sprint Prod. ????

Entity: Software
Attribute: Sprint
Unit: Size, Effort
Range: [0, ∞) [0, 1, 2, ..., n]

Productivity = Output / Input
## Process Measure – Sprint Productivity

**Name:** ’Average Sprint Productivity’ (~ ’Process Efficiency’)

<table>
<thead>
<tr>
<th>Name:</th>
<th>Avg. Sprint Prod.</th>
<th>Sprint Count</th>
<th>Sprint Output</th>
<th>Sprint Input</th>
</tr>
</thead>
</table>

**Entity:** Dev. Process

**Attribute:** Size

**Unit:**

**Range:**

\[
\text{\(\cdot\)} = TM \cdot TD \cdot WHD
\]

\[
\text{Range: } [0, \infty)
\]
Process Measure – Sprint Productivity

Name: ’Average Sprint Productivity’ (~ ’Process Efficiency’)

<table>
<thead>
<tr>
<th>Name</th>
<th>Avg. Sprint Prod.</th>
<th>Sprint Count</th>
<th>Sprint Output</th>
<th>Sprint Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>Efficiency</td>
<td>Sprint count</td>
<td>Size</td>
<td>Effort</td>
</tr>
<tr>
<td>Unit</td>
<td>LOC/person-hour</td>
<td>dimensionless</td>
<td>LOC</td>
<td>person-hour</td>
</tr>
<tr>
<td>Range</td>
<td>[0, ∞)</td>
<td>[0, 1, 2, ..., n]</td>
<td>[0, ∞)</td>
<td>[0, ∞)</td>
</tr>
</tbody>
</table>

Formula (Model): \( \text{ASP} [\text{loc/ph}] = \frac{\text{SUM}_{\text{sprint}} \left( \text{SO} [\text{loc}] / \text{SI} [\text{ph}] \right)}{\text{SC} [-]} \)

– Question: Does this formula make sense?

Note: Since a sprint has a fixed time in days (TD) and a fixed number of fulltime team-members (TM) with a fixed number of work-hours per day (WHD), one could ’Sprint Input’ (SI) consider a derived measure, calculated as \( \text{SI} [\text{ph}] = \text{TM} [\text{p}] \times \text{TD} [\text{d}] \times \text{WHD} [\text{h/d}] \)
The Goal / Question / Measure (Metric) Approach
The GQM Process

Business Goals

Initiate Measurement Project / Set Goals

Measurement Goals (Why ?)

Define Measures

GQM Plan (What ?)

Define Measurement (and Analysis)

Measurement Plan (How ?)

Analysis & Feedback

Execute Measurement (Collect data)

Measurement Data

Development Process

LTAT.05.008 / Software Analytics / © Dietmar Pfahl 2018
GQM Core Elements

GQM has three elements:

- Goals
- Questions
- Measures
GQM Core Elements: Goals

- GQM goal (or: Measurement Goals) are derived from business or improvement goals.

A GQM goal defines:
- which object is measured,
- for which purpose,
- with respect to which quality focus (aspect),
- from which viewpoint,
- and in which context (environment).

### GQM Goal Template

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
<td>What is analyzed?</td>
<td>Process, Product, Resource</td>
</tr>
<tr>
<td>Purpose</td>
<td>Why is the object analyzed?</td>
<td>Characterization, Monitoring, Improvement, ...</td>
</tr>
<tr>
<td>Quality Focus</td>
<td>Which characteristic of the object is analyzed?</td>
<td>Reliability, Flexibility, Maintainability, ...</td>
</tr>
<tr>
<td>Viewpoint</td>
<td>From which viewpoint is the quality focus analyzed?</td>
<td>Developer, Manager, Tester, Project Leader, ...</td>
</tr>
<tr>
<td>Context</td>
<td>In which context does the analysis take place?</td>
<td>Organization, Project, Application, ...</td>
</tr>
</tbody>
</table>
GQM Goal – Example

Analyze
test process
for the purpose of
characterization (understanding)
with respect to (quality aspect)
effectiveness
from the viewpoint of the
test team
in the environment of
project X, organization Y.
### GQM Question – Examples

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
<td>What is analyzed?</td>
<td>Process, Test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Product, Resource</td>
</tr>
<tr>
<td>Purpose</td>
<td>Why is the object analyzed?</td>
<td>Characterization,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Monitoring, Improvement,</td>
</tr>
<tr>
<td>Quality</td>
<td>Which characteristic of the object is analyzed?</td>
<td>Effectiveness,</td>
</tr>
<tr>
<td>Focus</td>
<td></td>
<td>Flexibility, Maintainability,</td>
</tr>
<tr>
<td>Viewpoint</td>
<td>From which viewpoint is the quality focus analyzed?</td>
<td>Developer, Manager, Tester, Project Leader,</td>
</tr>
<tr>
<td>Context</td>
<td>In which context does the analysis take place?</td>
<td>Organization, Project, Application,</td>
</tr>
</tbody>
</table>

- **Goal:** Analyze the *test process* for the purpose of *characterization* with respect to (quality aspect) *effectiveness* from the viewpoint of the test team in the environment of project X, organization Y.

- **Question 1:** How many failures are detected during testing?
- **Question 2:** When are failures detected (time)?
- **Question 3:** What types of failures are detected?
- **Question 4:** How much testing effort is spent?
- **Question 5:** Which test techniques/tools are applied?
- Etc.
Developing the GQM Hierarchy

Example GQM Hierarchy (incomplete):

- **Question 3**: What is the distribution of failures by criticality?
- **Model**: $D = F(x, y) = \frac{x[y]}{x[all]}$, $x = \text{Measure 1.1}$, $y = \text{Measure 3.1}$, where $D$: distribution of # failures per criticality class
- **Measure 1.1**: Failure count (ST: absolute; U: n/a; S: positive integer; O: product version 1.0)
- **Measure 3.1**: Failure criticality (ST: nominal; U: n/a; S: \{critical = complete breakdown of system, uncritical = unable to perform one or more of the functions F1, ..., F6, other\}, O: failure report

ST = Scale Type / U = Unit / S = Scale / O = Object (Entity)
GQM Plan

• The models and measures are defined by answering "What kind of information do we need in order to answer the questions?"
• The GQM-tree is documented in tabular form
• Each measure is defined by:
  – Name, ID
  – Scale, unit, etc.
  – Hypotheses
# Measurement Plan – Example

- Table for tracing Measurement Plan entries to GQM Plan, Project Plan (based on Process Model) and Data Collection Forms

<table>
<thead>
<tr>
<th>Goal-ID</th>
<th>Metric-ID</th>
<th>Metric-Name</th>
<th>Data Creation Event</th>
<th>Data Col. Time</th>
<th>Data Col. Resource</th>
<th>Data Provider</th>
<th>Data Collector</th>
<th>Form-Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Goal 1</td>
<td>M1.1</td>
<td>Failure count</td>
<td>Failure Report Summary</td>
<td>Test COMPLETE</td>
<td>TOOL: Failure Management System</td>
<td>Tester</td>
<td>QA Manager</td>
<td>Form X</td>
</tr>
<tr>
<td>Goal 1</td>
<td>M3.1</td>
<td>Failure criticality</td>
<td>Failure Report</td>
<td>Test report COMPLETE</td>
<td>TOOL: Failure Management System</td>
<td>Tester</td>
<td>QA Manager</td>
<td>Form X</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Goal 1</td>
<td>M4.1</td>
<td>Dev. team experience</td>
<td>Project team assignment</td>
<td>Project START</td>
<td>HUMAN: Interview or Questionnaire</td>
<td>Team member</td>
<td>Project Manager</td>
<td>Form Y</td>
</tr>
<tr>
<td>Goal 1</td>
<td>M5.1</td>
<td>Document count</td>
<td>CM system report</td>
<td>Test COMPLETE</td>
<td>TOOL: CM system</td>
<td>Developer / Tester</td>
<td>Project Manager</td>
<td>Form Z</td>
</tr>
<tr>
<td>Goal 1</td>
<td>M5.2</td>
<td>Document type</td>
<td>Document complete</td>
<td>Test COMPLETE</td>
<td>TOOL: CM system</td>
<td>Developer / Tester</td>
<td>Project Manager</td>
<td>Form Z</td>
</tr>
</tbody>
</table>
Data Collection

Principles
• Use multiple sources of data
• Create a case study database
• Validate data and maintain a chain of evidence

Methods
• direct (e.g., interviews)
• indirect (e.g., tool instrumentation)
• independent (e.g., documentation analysis)
Case Selection

- **Extreme/deviant** -- To obtain information on unusual cases, which can be especially problematic or especially good in a more closely defined sense.
- **Maximum variation** -- To obtain information about the significance of various circumstances for case process and outcome (e.g., three to four cases that are very different on one dimension: size, form of organization, location, budget).
- **Critical** -- To achieve information that permits logical deductions of the type, “If this is (not) valid for this case, then it applies to all (no) cases.”
- **Paradigmatic** -- To develop a metaphor or establish a school for the domain that the case concerns.

---

- In SE: mostly opportunistic, i.e., based on availability
Case Study Protocol

• See next slides
<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Background</td>
<td>(a) Identify previous research on the topic.</td>
</tr>
<tr>
<td></td>
<td>(b) Define the main research question being addressed by this study.</td>
</tr>
<tr>
<td></td>
<td>(c) Identify any additional research questions that will be addressed.</td>
</tr>
<tr>
<td>2. Design</td>
<td>(a) Identify whether single case or multiple case and embedded or holistic designs will be used, and show the logical links between these and the research questions.</td>
</tr>
<tr>
<td></td>
<td>(b) Describe the object of study (e.g., a new testing procedure; a new feature in a browser).</td>
</tr>
<tr>
<td></td>
<td>(c) Identify any propositions or subquestions derived from each research question and the measures to be used to investigate the propositions.</td>
</tr>
<tr>
<td>4. Procedures</td>
<td>(a) Procedures governing field procedures.</td>
</tr>
<tr>
<td>and roles</td>
<td>(b) Roles of case study research team members.</td>
</tr>
<tr>
<td>5. Data</td>
<td>(a) Identify the data to be collected.</td>
</tr>
<tr>
<td>collection</td>
<td>(b) Define a data collection plan.</td>
</tr>
<tr>
<td></td>
<td>(c) Define how the data will be stored.</td>
</tr>
<tr>
<td>Section</td>
<td>Content</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>6. Analysis</td>
<td>(a) Identify the criteria for interpreting case study findings.</td>
</tr>
<tr>
<td></td>
<td>(b) Identify which data elements are used to address which research question/subquestion/proposition and how the data elements will be combined to answer the question.</td>
</tr>
<tr>
<td></td>
<td>(c) Consider the range of possible outcomes and identify alternative explanations of the outcomes, and identify any information that is needed to distinguish between these.</td>
</tr>
<tr>
<td></td>
<td>(d) The analysis should take place as the case study task progresses.</td>
</tr>
<tr>
<td>7. Plan validity</td>
<td>(a) General: check plan against Höst and Runeson’s [73] checklist items for the design and the data collection plan (also in Appendix A of this book).</td>
</tr>
<tr>
<td></td>
<td>(b) Construct validity—show that the correct operational measures are</td>
</tr>
<tr>
<td>Section</td>
<td>Content</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>8. Study limitations</td>
<td>(c) Internal validity—show a causal relationship between outcomes and intervention/treatment (for explanatory or causal studies only). (d) External validity—identify the domain to which study finding can be generalized. Tactics include using theory for single-case studies and using multiple-case studies to investigate outcomes in different contexts. Specify residual validity issues including potential conflicts of interest (i.e., issues that are inherent in the problem, rather than arising from the plan).</td>
</tr>
<tr>
<td>9. Reporting</td>
<td>Identify target audience, relationship to larger studies</td>
</tr>
<tr>
<td>10. Schedule</td>
<td>Give time estimates for all of the major steps: planning, data collection, data analysis, reporting. Note data collection and data analysis are not expected to be sequential stages.</td>
</tr>
<tr>
<td>11. Appendices</td>
<td>(a) Validation: report results of checking plan against Höst and Runeson’s [73] checklist items (also in Appendix A of this book). (b) Divergences: update while conducting the study by noting any divergences from the above steps.</td>
</tr>
</tbody>
</table>
Case Study Checklist (for Review)

A.1 DESIGN OF THE CASE STUDY

1. What is the case and its units of analysis?
2. Are clear objectives, preliminary research questions, hypotheses (if any) defined in advance?
3. Is the theoretical basis—relation to existing literature or other cases—defined?
4. Are the authors' intentions with the research made clear?
5. Is the case adequately defined (size, domain, process, subjects)?
6. Is a cause–effect relation under study? If yes, is it possible to distinguish the cause from other factors using the proposed design?
7. Does the design involve data from multiple sources (data triangulation), using multiple methods (method triangulation)?
8. Is there a rationale behind the selection of subjects, roles, artifacts, viewpoints, and so on?
9. Is the specified case relevant to validly address the research questions (construct validity)?
10. Is the integrity of individuals/organizations taken into account?
Case Study Checklist (for Review)

A.2 DATA COLLECTION

11. Is a case study protocol for data collection and analysis derived (what, why, how, when)? Are procedures for its update defined?
12. Are multiple data sources and collection methods planned (triangulation)?
13. Are measurement instruments and procedures well defined (measurement definitions, interview questions)?
14. Are the planned methods and measurements sufficient to fulfill the objective of the study?
15. Is the study design approved by a review board, and has informed consent obtained from individuals and organizations?
16. Is data collected according to the case study protocol?
17. Is the observed phenomenon correctly implemented (e.g., to what extent is a design method under study actually used)?
18. Is data recorded to enable further analysis?
19. Are sensitive results identified (for individuals, the organization or the project)?
20. Are the data collection procedures well traceable?
21. Does the collected data provide ability to address the research question?
Case Study Checklist (for Review)

A.3 DATA ANALYSIS AND INTERPRETATION

22. Is the analysis methodology defined, including roles and review procedures?
23. Is a chain of evidence shown with traceable inferences from data to research questions and existing theory?
24. Are alternative perspectives and explanations used in the analysis?
25. Is a cause–effect relation under study? If yes, is it possible to distinguish the cause from other factors in the analysis?
26. Are there clear conclusions from the analysis, including recommendations for practice/further research?
27. Are threats to the validity analyzed in a systematic way and countermeasures taken? (Construct, internal, external, reliability)
Case Study Checklist (for Review)

A.4 REPORTING AND DISSEMINATION

28. Are the case and its units of analysis adequately presented?
29. Are the objective, the research questions and corresponding answers reported?
30. Are related theory and hypotheses clearly reported?
31. Are the data collection procedures presented, with relevant motivation?
32. Is sufficient raw data presented (e.g., real-life examples, quotations)?
33. Are the analysis procedures clearly reported?
34. Are threats to validity analyses reported along with countermeasures taken to reduce threats?
35. Are ethical issues reported openly (personal intentions, integrity issues, confidentiality)
36. Does the report contain conclusions, implications for practice, and future research?
37. Does the report give a realistic and credible impression?
38. Is the report suitable for its audience, easy to read, and well structured?
Case Study Checklist (for Review)

Summarizes the previous lists A.1-A.4

### A.5 READER’S CHECKLIST

39. Are the objective, research questions, and hypotheses (if applicable) clear and relevant? 1, 2, 5, 29, 30
40. Are the case and its units of analysis well defined? 1, 5, 28
41. Is the suitability of the case to address the research questions clearly motivated? 8, 9, 14
42. Is the case study based on theory or linked to existing literature? 3
43. Are the data collection procedures sufficient for the purpose of the case study (data sources, collection, validation)? 11, 13, 16, 18, 21, 31
44. Is sufficient raw data presented to provide understanding of the case and the analysis? 32
45. Are the analysis procedures sufficient for the purpose of the case study (repeatable, transparent)? 22, 33
46. Is a clear chain of evidence established from observations to conclusions? 6, 17, 20, 23, 25
47. Are threats to validity analyses conducted in a systematic way and are countermeasures taken to reduce threats? 27, 34, 37
48. Is triangulation applied (multiple collection and analysis methods, multiple authors, multiple theories)? 7, 12, 22, 24
49. Are ethical issues properly addressed (personal intentions, integrity, confidentiality, consent, review board approval)? 4, 10, 15, 19, 35
50. Are conclusions, implications for practice and future research, suitably reported for its audience? 26, 29, 36, 37, 38
Data Collection Methods

• Focus Groups

• Interviews
  – Unstructured
  – Semi-structured
  – Fully structured

• Observation

• Archival Data
  – Project repositories (open source & closed source)
  – Event logs, …
Data Collection Methods

• Focus Groups
  In a focus group, data collection is conducted with several people at the same time in a session resembling an interview

• Interviews
  – Unstructured
  – Semi-structured
  – Fully structured
  In interview-based data collection, the researcher normally asks a series of questions to one person.

• Archival Data
  – Project repositories (open source & closed source)
## Interview Types

<table>
<thead>
<tr>
<th></th>
<th>Unstructured</th>
<th>Semistructured</th>
<th>Fully Structured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical foci</td>
<td>How individuals qualitatively experience the phenomenon</td>
<td>How individuals qualitatively and quantitatively experience the phenomenon</td>
<td>Researcher seeks to find relations between constructs</td>
</tr>
<tr>
<td>Interview questions</td>
<td>Interview guide with areas to focus on</td>
<td>Mix of open and closed questions</td>
<td>Closed questions</td>
</tr>
<tr>
<td>Objective</td>
<td>Exploratory</td>
<td>Descriptive and explanatory</td>
<td>Descriptive and explanatory</td>
</tr>
</tbody>
</table>
Studies exploiting Archival Data in SE: Application Examples

• Journal: EMSE’16
  http://www.springer.com/computer/swe/journal/10664

• Conferences:
  – MSR’16: http://thomas-zimmermann.com/2016/01/msr-2016/
  – ESEM’16: http://alarcos.esi.uclm.es/eseiw2016/esem
  – EASE’16: http://ease2016.lero.ie
  – PROMISE’16: http://promisedata.org/2016/
SE Data Repositories

• App stores (Google Play, etc.)
• Q/A web-pages (e.g., StackOverflow)
• Crash report repositories (e.g., Ubuntu’s repository)
• YouTube tutorials (e.g., tool tutorials)
• Industry data: ISBSG repository, Finnish dataset
• Issue Trackers -> e.g. JIRA
• Version Control Systems -> e.g., Git
• Tera-PROMISE: http://openscience.us/repo/
Welcome to one of the largest repositories of SE research data

The tera-PROMISE Repository is a research dataset repository specializing in software engineering research datasets. We offer free and long-term storage for your research artifacts. Learn more on our about page.

How to Reference Us:


You can view all of our datasets in the categories listed on the left and on the categories page.

Find research datasets

We have everything from McCabe & Halsted to Spreadsheets to Green Mining.

Contribute your data

Learn how to contribute your research data, whether you’re a researcher or a student.

View categories

Learn how
GitHub & GHTorrent

- [https://en.wikipedia.org/wiki/GitHub](https://en.wikipedia.org/wiki/GitHub)
- API’s for Java, Ruby, Python, etc.

Articles:

“The GHTorrent Dataset and Tool Suite” (2013)

“Lean GHTorrent: GitHub data on demand“ (2014)
GHTorrent

<table>
<thead>
<tr>
<th>Entity</th>
<th>Description</th>
<th>Raw data entity</th>
<th>Num Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>__</td>
<td></td>
<td>__</td>
<td>__</td>
</tr>
<tr>
<td>projects</td>
<td>Project repositories.</td>
<td>repos</td>
<td>1,326,900</td>
</tr>
<tr>
<td>users</td>
<td>Github users.</td>
<td>users</td>
<td>793,855</td>
</tr>
<tr>
<td>project_members</td>
<td>Users with commit access to the referenced project.</td>
<td>repo_collabs</td>
<td>983,629</td>
</tr>
<tr>
<td>organization_members</td>
<td>List of members in an organization.</td>
<td>org_members</td>
<td>34,924</td>
</tr>
<tr>
<td>commits</td>
<td>A list of all commits on Github. The project_id field refers to the first project this commit has been added to.</td>
<td>commits</td>
<td>29,978,291</td>
</tr>
<tr>
<td>project_commits</td>
<td>List of all commits to a project.</td>
<td>__</td>
<td>__</td>
</tr>
<tr>
<td>commit_parents</td>
<td>Commits that are parents to a commit.</td>
<td>__</td>
<td>__</td>
</tr>
<tr>
<td>commit_comments</td>
<td>Code review comments for a commit.</td>
<td>commit_comments</td>
<td>126,697</td>
</tr>
<tr>
<td>watchers</td>
<td>users that have starred (was watched) a project</td>
<td>watchers</td>
<td>7,744,619</td>
</tr>
<tr>
<td>followers</td>
<td>users that are following another user.</td>
<td>followers</td>
<td>1,797,343</td>
</tr>
<tr>
<td>issues</td>
<td>Issues that have been recorded for a project.</td>
<td>issues</td>
<td>2,326,069</td>
</tr>
<tr>
<td>issue_events</td>
<td>Chronologically ordered list of events on an issue.</td>
<td>issue_events</td>
<td>4,085,294</td>
</tr>
<tr>
<td>issue_comments</td>
<td>Discussion comments on an issue.</td>
<td>issue_comments</td>
<td>2,886,006</td>
</tr>
<tr>
<td>pull_requests</td>
<td>List of pull requests for base_repo. Requests originate at head head_repo/commit and are created by user_id</td>
<td>pull_requests</td>
<td>1,144,251</td>
</tr>
<tr>
<td>pull_request_comments</td>
<td>Discussion comments on a pull_request</td>
<td>pullreq_comnts</td>
<td>2,228,894</td>
</tr>
<tr>
<td>pull_request_history</td>
<td>Chronologically ordered list of events on a pull_request</td>
<td>__</td>
<td>__</td>
</tr>
</tbody>
</table>

Fig. 1. Schema entities, their description, the corresponding raw data entities and the number of raw data items (Feb 15, 2013).
GHTorrent – Database Dumps

http://ghtorrent.org/downloads.html
Data Analysis – Qualitative Data

Main steps of analysis

Data Analysis

1. Data collection
2. Coding
3. Hypothesis definition
4. Generalization / findings
5. Reporting
Data Analysis – Quantitative Data

• Descriptive statistics
  – Examples: mean values, standard deviations, histograms, scatter plots
  – Used to get an understanding of the data that has been collected
  – Often a natural step before any other methods are applied

• Development of predictive models
  – Conducted in order to describe how a measurement from a later process activity is related to an earlier process measurement

• Hypothesis testing
  – Conducted in order to determine if there is a significant effect of one or several variables (independent variables) on one or several other variables (dependent variables)
Threats to Validity

• Construct Validity
  – adequate instruments/constructs/measures adequate?

• Internal Validity
  – adequate design? (relates to causality / confounding factors)

• Conclusion Validity
  – adequate statistical approaches?

• External Validity
  – results generalizable?

• Reliability
  – results robust/repeatable?
# Case Study Reporting Structure

<table>
<thead>
<tr>
<th>Section headings</th>
<th>Subsections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td></td>
</tr>
<tr>
<td>Authorship</td>
<td></td>
</tr>
<tr>
<td>Structured abstract</td>
<td>context – goal – method – result – conclusion</td>
</tr>
<tr>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>Problem statement / motivation</td>
<td></td>
</tr>
<tr>
<td>Research objectives</td>
<td></td>
</tr>
<tr>
<td>Context</td>
<td></td>
</tr>
<tr>
<td>Related work</td>
<td></td>
</tr>
<tr>
<td>Earlier studies</td>
<td></td>
</tr>
<tr>
<td>Theory</td>
<td></td>
</tr>
<tr>
<td>Case study design</td>
<td></td>
</tr>
<tr>
<td>Research questions</td>
<td></td>
</tr>
<tr>
<td>Case and subject selection</td>
<td></td>
</tr>
<tr>
<td>Data collection procedure(s)</td>
<td></td>
</tr>
<tr>
<td>Analysis procedure(s)</td>
<td></td>
</tr>
<tr>
<td>Validity/evaluation procedure(s)</td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td></td>
</tr>
<tr>
<td>Case and subject descriptions, covering execution, analysis and implementation issues</td>
<td></td>
</tr>
<tr>
<td>Subsections, which may be structured, for example, according to coding scheme, each linking observations to conclusions Evaluation of validity</td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td></td>
</tr>
<tr>
<td>Impact/implications</td>
<td></td>
</tr>
<tr>
<td>Relation to existing evidence</td>
<td></td>
</tr>
<tr>
<td>Limitation</td>
<td></td>
</tr>
<tr>
<td>Conclusions &amp; future work</td>
<td></td>
</tr>
<tr>
<td>Summary of findings</td>
<td></td>
</tr>
<tr>
<td>Future work</td>
<td></td>
</tr>
<tr>
<td>Acknowledgements</td>
<td></td>
</tr>
<tr>
<td>References</td>
<td></td>
</tr>
<tr>
<td>Appendices</td>
<td></td>
</tr>
</tbody>
</table>
Relevant Literature (Selection)


Homework 1: Solution
Homework 2: Case Study Research