Pre-processing of log-data: Adapting familiar routines to unfamiliar data

Frankfurt, June 17

Heiko Sibberns & Ulf Kroehne

IEA

Researching education, improving learning

DIPF

Leibniz Institute for Research and Information in Education

CENTRE FOR INTERNATIONAL STUDENT ASSESSMENT

Content

- Pre-processing in the analyses workflow
- Common data and log-data
- Events – General and specific attributes
- Preparation and parsing
- Generation of datasets (Transformation)
- Purifying and anonymizing log-data

Workflow: Pre-Processing->Feature Extraction->Analyses and Validation

Phase	Steps and Purpose	Term
Pre-Processing	- Preparing and parsing log-data (cleaning)	Log-Data ↓
	- Generation of datasets (transformation)	
	- Checking and purifying log data (including anonymization)	
	- Documentation of log-data	
Feature Extraction	- Validation of log-data (sequences, contexts)	↓ Process Indicators
	- Theoretical derivation of indicators	
	- Extraction of features and indicators	
	- Combination of low-level features to process indicators	
Analysis and Validation	- Documentation of indicators	
	- Life cycle of process indicators validation	

Pre-Processing

Phase	Steps and Purpose	Term
Pre-Processing	- Preparing and parsing log-data (cleaning)	Log-Data ↓
	- Generation of datasets (transformation)	
	- Checking and purifying log data (including anonymization)	
	- Documentation of log data	

- **Focus will be on pre-processing step. Documentation is a separate topic which will be addressed by Britt He.**

Common data in ILSAs: Example from TIMSS 2015 DC

	IDCNTRY	IDSCHOOL	ACBG03A	ACBG03B	ACBG04	ACBG05A	ACBG05B	ACBG06A	ACBG06B	ACBG07A	ACBG07B	ACBG07C	ACBG0
1	276	1	2	1	3	6	4	3	3	2	3	2	
2	276	2	1	3	1	4	4	3	3	3	2	2	
3	276	3	3	1	5	2	3	3	3	3	3	2	
4	276	4	4	1	4	2	1	3	2	2	2	2	
5	276	5	2	2	2	2	2	2	3	1	2	2	
6	276	6	2	1	2	2	2	3	2	1	1	1	
7	276	7	2	1	4	6	4	3	3	3	2	3	
8	276	8	2	2	2	5	3	3	3	2	2	2	
9	276	9	4	2	4	5	3	3	3	1	2	2	
10	276	10	1	2	5	6	4	3	3	2	2	3	
11	276	11	2	1	3	6	5	3	3	2	2	1	
12	276	12	2	3	2	2	2	3	3	1	4	2	
13	276	13	3	2	3	4	3	3	2	2	2	2	
14	276	14	4	1	5	2	2	3	3	1	1	1	
15	276	15	4	1	2	2	2	3	3	1	1	1	
16	276	16	4	1	4	1	1	3	3	1	3	2	
17	276	17	.	.	1	6	4	3	3	1	3	1	
18	276	18	A	A	A	A	A	A	A	A	A	A	
19	276	19	A	A	A	A	A	A	A	A	A	A	
20	276	20	.	.	1	7	4	3	3	1	3	1	
21	276	21	.	.	1	7	5	3	3	1	2	2	
22	276	22	2	1	3	6	4	3	3	2	2	2	
23	276	23	A	A	A	A	A	A	A	A	A	A	
24	276	24	A	A	A	A	A	A	A	A	A	A	
25	276	25			1	7	4	3	3	2	2	2	

Common data

- Clear structure
- Matrix –format
- One line per respondent, variables in columns
- Meta-data are availale in codebooks
- Tables in relational databases, 1:n matches that allow to link schools-classes-students
- Since decades THE data structure in ILSAs
- Common formats are SPSS, SAS, STATA, (R)....

Log-file data: Example from digitalPIRLS FT


```
riable_id":"ExtendedTextInteraction-D021O20C","duration":"104","attempted":"false","answer":"","time
duration":"16","attempted":"true","answer":"D021V07B","timestamp":"2020-03-11 09:09:29.887","is_fina
020-03-11 09:23:26.664","is_final_answer":"true"},{"learner_response_id":99814637,"variable_id":"Ext
nswer":"true"},{"learner_response_id":99815252,"variable_id":"ChoiceInteraction-SQ9","duration":"8",
BC ASXR06C:ASXR06CB ASXR06D:ASXR06DB ASXR06E:ASXR06EC ASXR06F:ASXR06FB ASXR06G:ASXR06GC","timestamp"
n ger\u00e4tetaucher.","timestamp":"2020-03-12 07:24:53.646","is_final_answer":"true"},{"learner_res
:"179","attempted":"true","answer":"Warum machen s sie dann nicht","timestamp":"2020-03-12 07:53:19.
:"ChoiceInteraction-D021V06","duration":"127","attempted":"true","answer":"D021V06B","timestamp":"20
","is_final_answer":"true"},{"learner_response_id":99879303,"variable_id":"InlineChoiceInteraction-D
ble_id":"ChoiceInteraction-SQ5b","duration":"14","attempted":"true","answer":"ASBG02BC","timestamp":
"true","answer":"ASBR04D","timestamp":"2020-03-12 09:40:14.529","is_final_answer":"false"},{"learner
R01AB ASBR01B:ASBR01BB ASBR01C:ASBR01CA ASBR01D:ASBR01DB ASBR01E:ASBR01EA","timestamp":"2020-03-12 0
is_final_answer":"true"},{"learner_response_id":99881587,"variable_id":"GapMatchInteraction-SQ2","du
99882036,"variable_id":"GapMatchInteraction-SQ18","duration":"0","attempted":"true","answer":"ASBR01
tion":"91","attempted":null,"answer":"","timestamp":"2020-03-12 07:09:19.842","is_final_answer":"tru
e_id":"ExtendedTextInteraction-D021O14A","duration":"3","attempted":"false","answer":"","timestamp":
3-12 07:51:42.428","is_final_answer":"false"},{"learner_response_id":99940222,"variable_id":"Extende
tory":[{"learner_response_id":99937800,"variable_id":"ChoiceInteraction-D021V01","duration":"1","att
2","attempted":"false","answer":"","timestamp":"2020-03-12 09:17:38.061","is_final_answer":"false"},
2047,"variable_id":"ExtendedTextInteraction-D021V06","duration":"134","attempted":"true","answer":"w
nse_id":99943442,"variable_id":"ChoiceInteraction-D021V12","duration":null,"attempted":"true","answe
nswer":"","timestamp":"2020-03-12 09:18:05.940","is_final_answer":"true"},{"learner_response_id":999
BG05IB ASXG05J:ASXG05JA","timestamp":"2020-03-12 09:31:19.304","is_final_answer":"true"},{"learner_r
```


IEA

Log-file data

- Formats are highly dependent on the system that generated them.
- At a first glance, somehow „messy“
- In the past, often collected for debugging purposes during programming
- Content can be dependent on needs and conventions in development team ...and sometimes personal „taste“ of individual developers.
- Often no formal description of data available (seconday use was not anticipated)
- However, this is changing since research needs are adressed more closely

IEA

Log-file data to capture events

- XML, JSON or CSV formats are common
- Relational and non-relational databases
- Specific formats for event stream data (XES)
- „Mixed“ formats are also used, e.g. JSON in a column of a SQL-table
- More formats from the e-learning community
- In a nutshell: Each event has one entry – one row if events are recorded in tabular format.
- Events are stored with one or more attributes

IEA

Minimal content of log-file Events – General Attributes

- Identifier linking the event to a specific respondent
- Link to a specific part or unit, page, item or index
- Timestamp, indication when the event was triggered or completed
  - Absolute or relative
  - ISO 8621
  - UNIX timestamp (time elapsed since Jan 1, 1970)
- Event type
- Event type-specific data, i.e. specific attributes

-> Raw log data

Additional /optional attributes

- Login-identifier to monitor for several log-ins, either intentional-e.g. in games- or due to system failures like program crashes or lost server connections
- More specific information on the page, e.g. if rotated test forms are used

Type-specific attributes

- Number and characteristic of attributes depend on the event-type
  - Event-type: BtnNext
  - Attributes: {"event":"clicked"}
  - One „trivial“ attribute is saved

  - Event-type: LogIn
  - Attributes: {"culture":"enUS","grade":4,"alphaIso":"USA","loginState":"success","windowHeight":912,"windowWidth":1368,"windowDevicePixelRatio":2,"userAgent":"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.98 Safari/537.36","currentDevice":3,"isSpecialAccommodated":0}

Decomposition of type-specific attributes

- Nine attributes are allocated to the event LogIn
- In general, attributes can be decomposed into variables
- Attributes in this example can be decomposed into nine variables, e.g. Culture, Grade, CountryALPHA,LoginState...
  - Non-missing values only for the event LogIn
  - All other events will show missing value for these variables->sparse data matrix

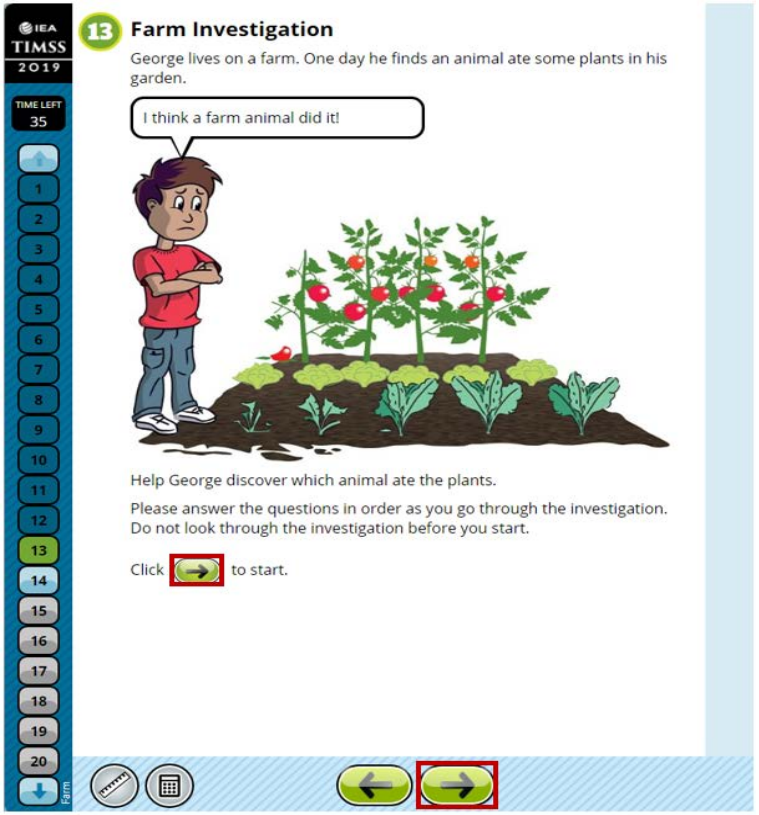
Pre-processing event-type specific attributes

- Parsing the string with attributes
- Disassemble attributes by event
- Challenge: Different events come with different attributes

Note on event-types and event-specific data

- Technical platform-specific documentation and specification of events and event-types is necessary (e.g., XSD –XML schema definition or JSON schema)
- Further standardization of event-types and event-specific data would be nice, but is this possible across platforms?
- Documentation of interactive computer-based instruments must provide detailed information to understand the events and what has been saved.
- Different formats are possible, e.g., interactive mock-items, html-documentation or a detailed description including screenshots

Documentation: Example from eTIMSS



aiuID	CurrentIndex	Event Attributes
13616	12	{"event":"clicked"}
13616	13	{"action":"goTo:item","from":{"blockName":"1","aiuId":13616,"index":12},"to":{ "blockName":"1","aiuId":13616,"index":13}}

Purpose of data cleaning

- Eliminate errors
- Increase data reliability
- Ensure consistency
- Run plausbility check to avoid conflicting interpretations (e.g., same final raw response in last log-event and result data)
- Assure completeness (to the extend possible)

General Cleaning of log-file data - examples

- Check for multiple identical uploads
- Check for invalid upload like test uploads
- Examination of multiple logins of a respondent due to system crashes or lost server connections. Decision, which data should be kept or combined
- Check for consistency with other data sources like tracking forms or sampling data
- Check for implausible timestamp information
  - Negative time intervals
  - Unreasonable time intervals, but what is „unreasonable“
- Checking for events, that were triggered by the delivery system and are caused by system failure.

Transformation: Representation of log-data

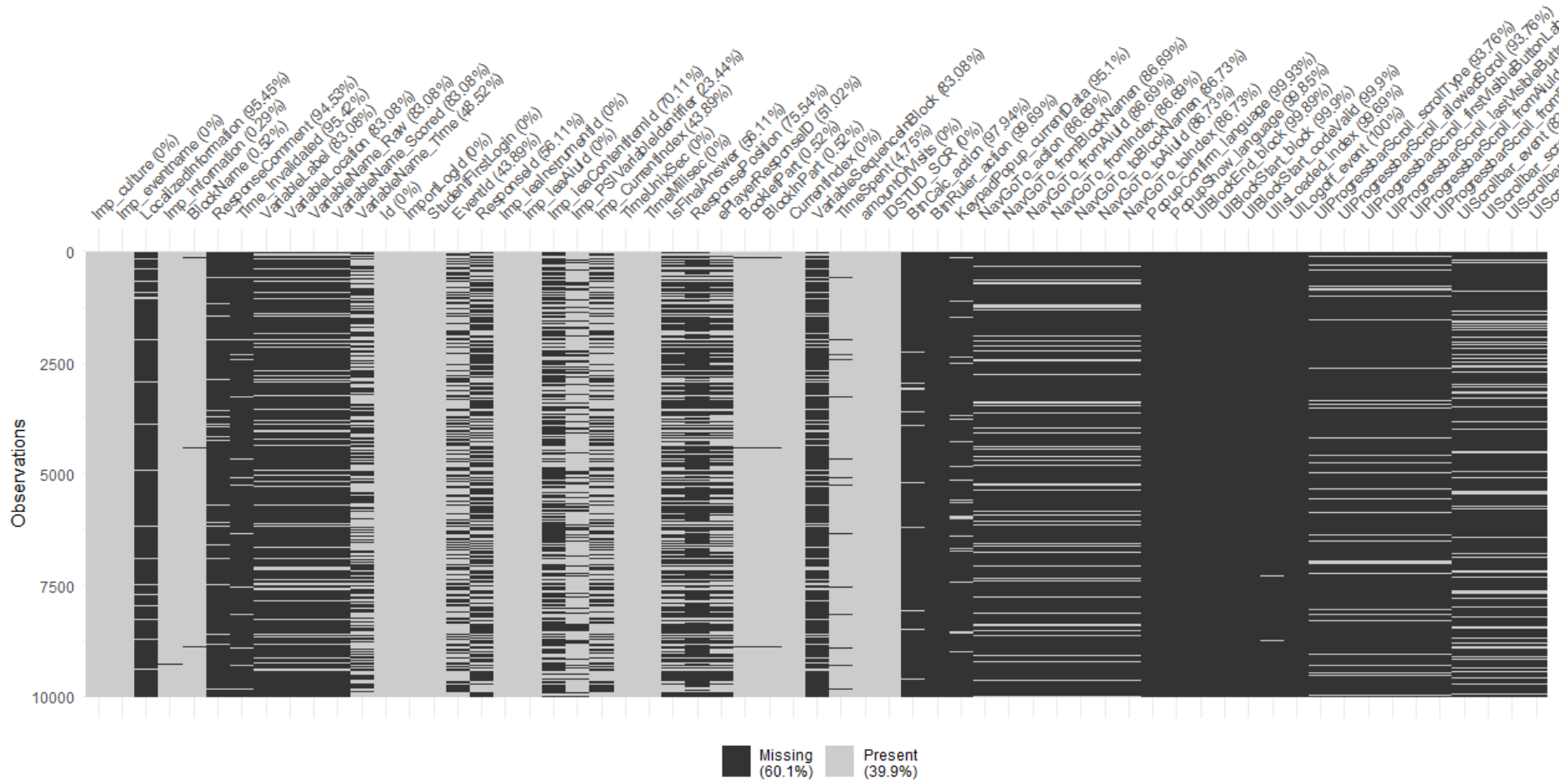
- **Flat and sparse log-data table**
  - Combined table containing all event data
  - Event-specific attributes/data in separate columns
  - Missing values for event-specific attributes NOT provided for an event of a particular type

- **Universal log-format**
  - One table for each event type
  - In each table event-specific attributes in columns
  - Missing values only for optional attributes

IEA

Flat and sparse log-data table

Option 1: Flat and sparse log-data table



Option 2: Universal log-format

log.BtnNext	
PersonIdentifier	numeric
EventName	character
TimeStamp	character
Element	numeric
Id	numeric
ImportLogId	numeric
Imp_culture	character
Imp_ieaInstrumentId	numeric
BlockName	character
BlockInPart	numeric
BookletPart	numeric
StudentFirstLogin	numeric
CurrentIndex	numeric
TimeSpent	numeric
amountOfVisits	numeric
Time_Invalidated	character
VariableName_Time	character
EventId	numeric
Imp_ieaContentItemId	numeric
Imp_PSIVariableIdentifier	numeric
Imp_CurrentIndex	numeric

Purifying data

- Delete redundant information
- Keep only events necessary for a specific analyses
- Integrity checks for correct sequence of events
- Anonymization
  - Scrambled or masked identifiers only
  - Removal of names or other characteristics that could help to identify a person from response strings

Thank you

Researching education, improving learning