An overview of plant data for machine learning: categories, availability, and common problems

Ruomu Tan*, Benjamin Klöpper

Motivation

- The report [1] gives an overview of the categories of data that are typically available in chemical plants.
- Plant data can be leveraged when developing machine learning model for operation support. The report also highlights the availability and typical problems when using the data.
- We designed and distributed a questionnaire [2] with use case owners to collect feedback based on the experience with real-life data from industrial plants.
- As a deliverable of TP5, the report will be useful for researchers, especially those with limited experience of real-life data, as a starting point for understanding and exploration of plant data before developing and deploying ML solutions.

Availability of each category of data

- Several examples of data availability are shown as follows:

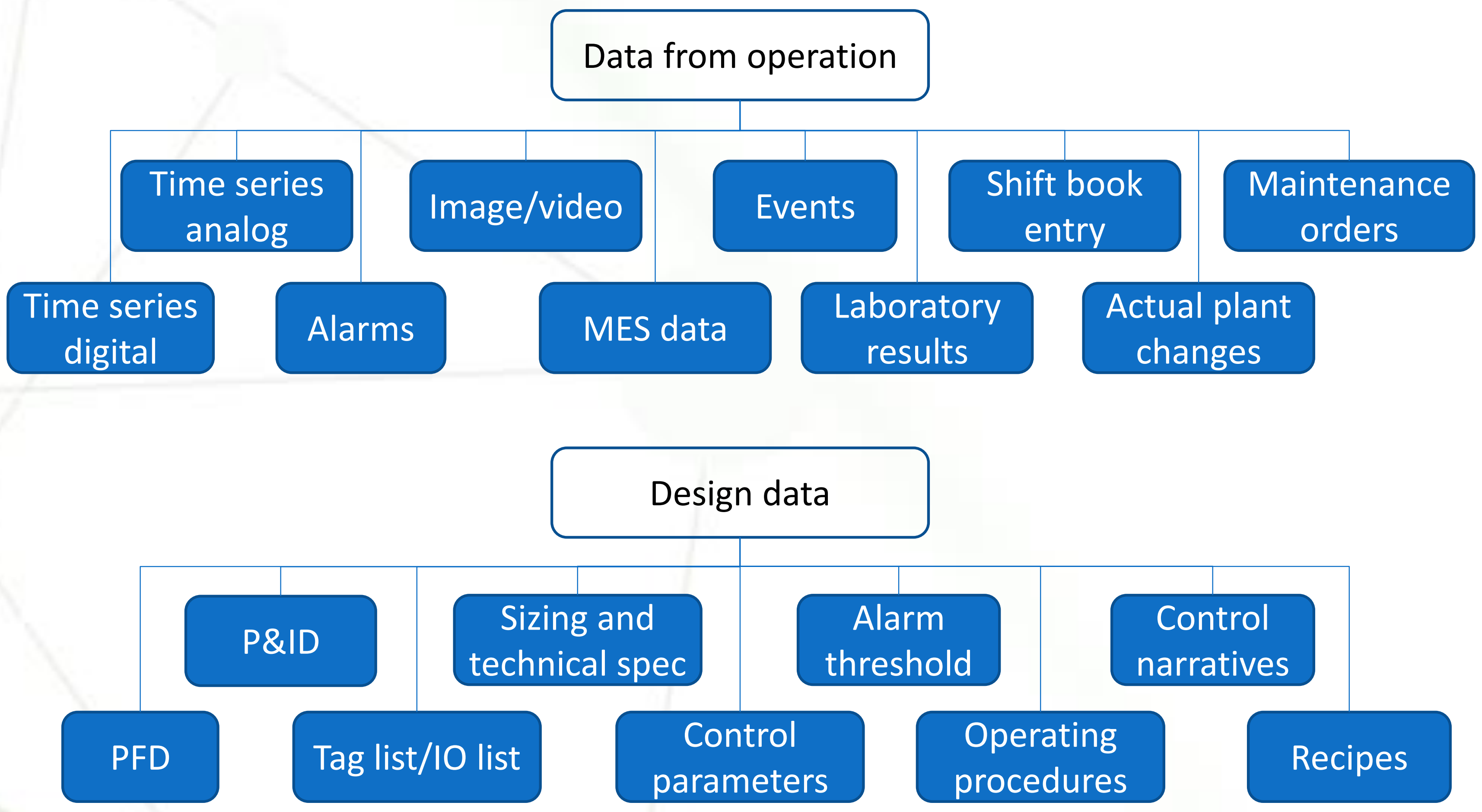
Data	Availability	Time window	For sharing?
Time series	(almost) Always	Several years	NDA needed
Alarm&Event	Sometimes events not available	Months to years	NDA needed
Shift-book entry	Available in free text	Several years	Strict NDA due to data protection
MES data	Available, more frequently seen in batch processes	A few years	NDA needed
Image/video	Rarely available	Often not stored	N/A
Design data	Always	N/A	Anonymization needed

- The data availability vary significantly among the categories. There may exist knowledge gaps in utilizing less available data for ML solutions.
- The restriction in data sharing may also impact the usage of data.

Conclusions and next steps

- Conclusions:
  - The availability of each category of data can differ greatly due to the characteristic of plants and the configuration of the data collection systems.
  - Some categories of data, e.g., time trends of process variables, are much more frequently used than the others.
  - Multiple data categories are connected and sometimes complementing one another.
  - Design data may always exist; however the bottleneck of using such data is anonymization.
- Next steps
  - Continuous effort to collect feedback and improving the report;
  - Exploration of the data categories that are less used;
  - Fusion of data from multiple categories.

Data categories



Common problems when using the data

- Two examples of the common problems in the data

Category	Problems
Laboratory results	- Time delay of the lab results - Sparseness and multiple sampling rate - Human error when doing lab analysis - Unreliable timestamps when recording the results - Mapping problem when a sample represents the accumulated status of the plant - Changes in measurements/reporting
Alarms and events (A&E)	- Data format is different from time series - Irrelevant A&E data for a certain purpose - Not reliable events when an alarm is acknowledged - Time delay between the event of PVs change and actual change in the PV trends - Incomplete A&E data when a relevant status change is not recorded

Acknowledgement

- The authors would like to thank the following KEEN colleagues:
  - Marco Gärtler, Silke Merkelbach and Valentin Khaydarov for distributing the questionnaire to the use case owners in TP5;
  - Franz Bähner, Supasuda Assawajaruwan and Laura Neuendorf for kindly providing inputs to the questionnaire;
  - Simone Rogg for supporting the activity and providing guidance for collecting feedback.
- We also looking forward to your inputs in the future!

* Presenter and contact person

[1] R. Tan and B. Kloepper: Plant data for machine learning – types of data, availability, interraltions, common problems and challenges. Technical report.

[2] Questionnaire for plant data description.

Both documents are available in: https://keen-austausch.de/apps/files/?dir=/TP5_Prozessdaten/AP5. 4/Data&fileid=6508

www.keen-plattform.de

Contact

Ruomu Tan

ABB Corporate Research Center

www.abb.com

ruomu.tan@de.abb.com

Supported by:

Federal Ministry for Economic Affairs and Energy

on the basis of a decision by the German Bundestag