Towards A Data Repository for Educational Factories

Fajar J. Ekaputra[1] Martin Weise[1] Katharina Flicker[1]
Mohd. Rizal bin Salleh[2] Md. Nizam Abd Rahman[2]
Azrul Azwan Abd Rahman[2] Tomasz Miksa[1] Andreas Rauber[1]

[1]Technische Universität Wien, Austria
[2]Universiti Teknikal Malaysia Melaka, Malaysia

ICoDSE, November 2nd 2022

Introduction

Current and future workforce needs to be equipped with digital world knowledge to:

- Ensure seamless transition of graduates, into a
- Working environment that expects such skill sets

Problem:

- Identify key requirements for students to learn these skill sets to use them for basic and applied research
- Heterogeneity of tools and machines in the modern factory setup to provide such data in reliable manner

Introduction

Proposal:

- Standardized data management and -analytics process for educational factories (e.g. Pilot Factory at TU Wien or Teaching Factory at UTeM)
- Specialized Database Repository
- Define the data capturing, provenance collection and data management approaches in these educational factories

Contribution:

- Ontologies to enrich data with semantics, ease the integration of advanced data analytics and exploration tools
- Expose students from various disciplines and faculties to various aspects of the digital world: data capturing process, equipment, etc.

Preliminaries

UTeM Teaching Factory:

- Provide real-life environment simulation of productions and services in everyday industrial practices
- Produce component and competitive graduates
- Students can discover and solve problems experimentally in a factory/industrial setting

Overall goal

Increase the chance of creating the creative thinkers for the current and future engineers and IT professionals

Preliminaries

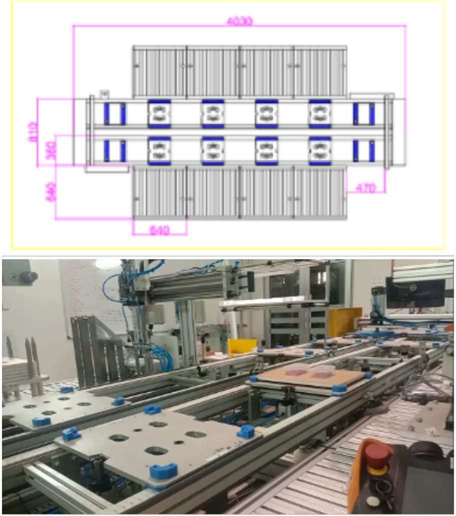


Figure: UTeM's computer-integrated manufacturing line

Preliminaries

Pilot Factory of TU Wien:

- Discrete, multi-variant series production, production on very small quantities
- Coverage of all phases, from design to assembly (targeted towards medium-sized companies)
- Develop scientific know-how on optimal production techniques, benefits the economy
- Crucial role in teaching at TU Wien in experimental setting

Overall goal

Support advanced training and research on how the future industry should look like, support researcher and students in conducting their studies and research.

Preliminaries

Figure: Pilot Factory of TU Wien

Preliminaries

Database Repository [8] for research data stored in databases:

- Allow researchers to deposit their research data from the start of a project in a private cloud setting usually only available in large-scale infrastructures
- Separation of concerns, letting researchers do research, while data stewards handle classic data management tasks
- Improving *findability*, *accessibility*, *interoperability* and *reusability* [10] of the metadata
- Allowing reproducability of arbitrary subsets of evolving data [6]

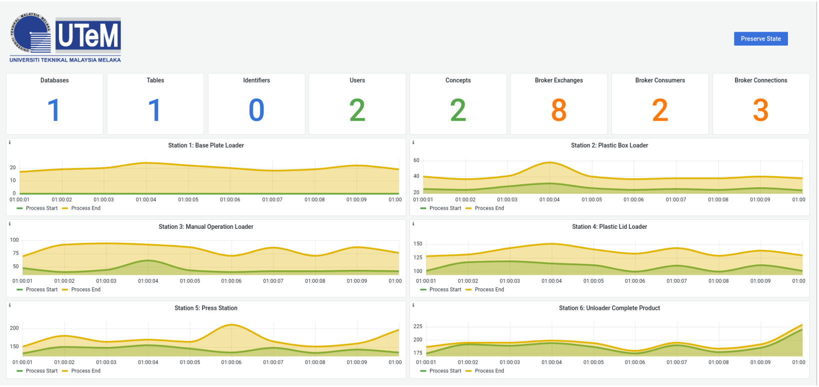
Proposed Approach

Core requirements

Series of discussions and interviews (TU Wien Pilot Factory and UTeM's Teaching Factory) and non-public instance for general research data management:

- Data capturing from machineries and sensors through PLC and push to a data sink (e.g. relational database) via open and documented interfaces with lightweight authentication,
- Dedicated hardware that is maintained by trained IT-personnel who will take care of administration tasks to ensure availability and security of the hardware for deployment and operation,
- Utilization of collected data through providing computational capabilities and visualization of the data to e.g. students through own software scripts

Proposed Approach



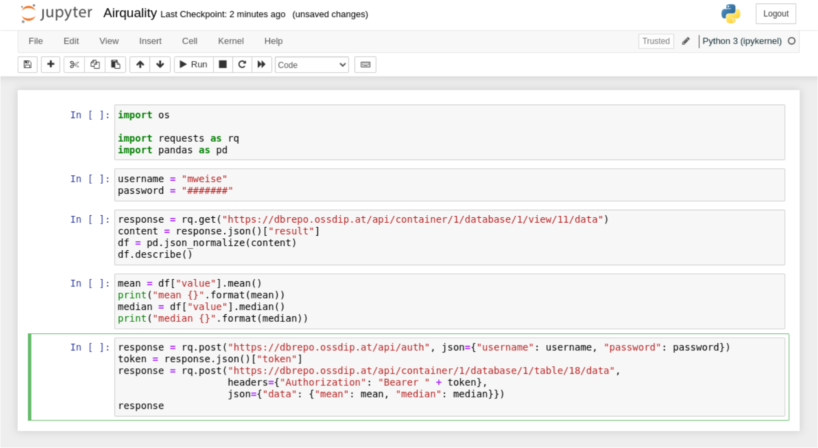
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Preserve State

Databases	Tables	Identifiers	Users	Concepts	Broker Exchanges	Broker Consumers	Broker Connections
1	1	0	2	2	8	2	3

Figure: Visualization of collected data as real-time visualization

Proposed Approach



```
In [ ]: import os

        import requests as rq
        import pandas as pd
```

```
In [ ]: username = "mweise"
        password = "#######"
```

```
In [ ]: response = rq.get("https://dbrepo.ossdip.at/api/container/1/database/1/view/11/data")
        content = response.json()["result"]
        df = pd.json_normalize(content)
        df.describe()
```

```
In [ ]: mean = df["value"].mean()
        print("mean {}".format(mean))
        median = df["value"].median()
        print("median {}".format(median))
```

```
In [ ]: response = rq.post("https://dbrepo.ossdip.at/api/auth", json={"username": username, "password": password})
        token = response.json()["token"]
        response = rq.post("https://dbrepo.ossdip.at/api/container/1/database/1/table/18/data",
                           headers={"Authorization": "Bearer " + token},
                           json={"data": {"mean": mean, "median": median}})
        response
```

Figure: Providing compute-capabilities

Proposed Approach

Architecture of 14 micro services who are responsible for an atomic task of operating the system as a whole:

- Encapsulate one service in one Docker container (providing necessary libraries and runtime configuration),
- Custom deployments and exchanging any service with modified implementations (e.g. organizational SSO complimentary to the provided JWT authentication),
- Managing the lifecycle of data through a pull-based API following REST [1] via HTTP, adding data via AMQP or connecting directly through JDBC (currently only by the services).

Proposed Approach

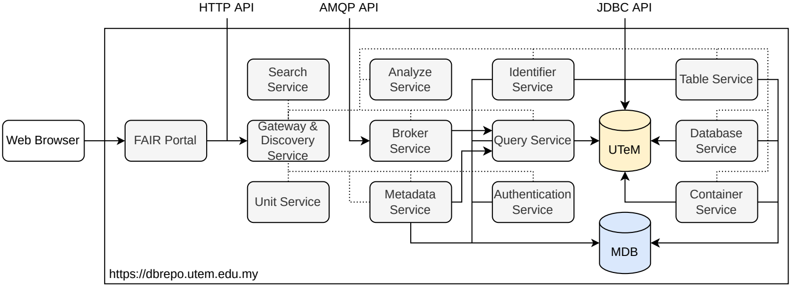
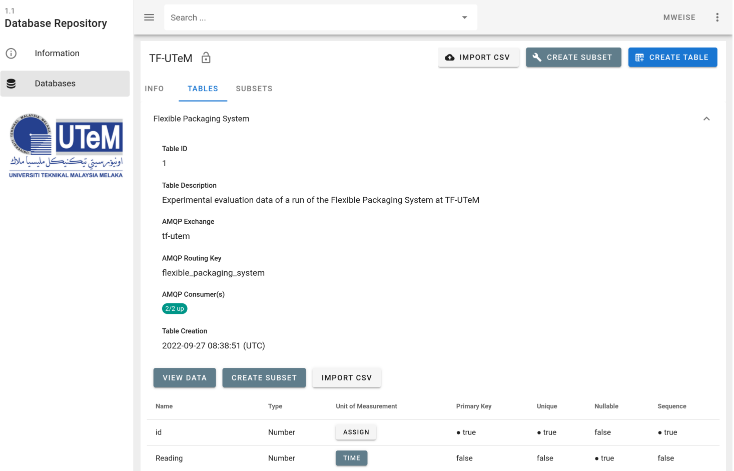


Figure: Micro service architecture of DBRepo offering three machine-actionable APIs and a GUI (FAIR Portal)

Proposed Approach



Flexible Packaging System

Table ID
1

Table Description
Experimental evaluation data of a run of the Flexible Packaging System at TF-UTeM

AMQP Exchange
tf-utem

AMQP Routing Key
flexible_packaging_system

AMQP Consumer(s)
2/2 up

Table Creation
2022-09-27 08:38:51 (UTC)

Name	Type	Unit of Measurement	Primary Key	Unique	Nullable	Sequence
id	Number	ASSIGN	● true	● true	false	● true
Reading	Number	TIME	false	false	● true	false

Figure: The FAIR Portal describes the data tables through metadata

Discussion

Curation of data requires both technical and organizational commitment from the organization that wants to make manufacturing data from educational factories available to students:

- Already performed for large-scale infrastructures for e.g. climate- or genome data
- Not in smaller infrastructures such as TU Wien Pilot Factory or UTeM's Teaching Factory
- Existing generalist repositories ease the start for data management but are not fit for continuously evolving data stored in e.g. databases

Discussion

Impact further depends on:

- Interoperability of PLCs and the repository through interfaces (e.g. HTTP API, AMQP API, JDBC API)
- Offer interaction capabilities to students: access roles for private/public databases, creation of views for e.g. embargoed access, addition/correction/deletion of data while keeping track of changes
- Necessary schema changes create a challenge and are subject to future work (due to the reproducability of queries at arbitrary points in time)

Discussion

Complex data infrastructures safeguarding data that is collected at a repository, possible after expensive (pre-)processing and ensuring their proper documentation and availability to researchers with high quality:

- Establishment and development of these specialized infrastructures bears risks of failure to address the specific needs of a research community
- Doubts of trustworthiness or data quality, researchers may refrain from using these infrastructures and data

Consitute trust in infrastructures and data quality

Elicit actual requirements, create trust in infrastructures and data quality through (i) data provenance (ii) establishment of mechanisms that enable verification of data quality, and (iii) transparency.

Discussion

Interviews reveal understanding of *transparency*:

- To understand data reuse, researchers want to have insight what data was reused, to what extent and to what success degree,
- Data access allowance bound to a role, transparency alone is insufficient

Trust in infrastructures and its components is essential for researchers to trust one's own research outputs and to accept the accountability when releasing research results into society

Related Work

Repositories in the literature are manifold, two types are identified: (1) discipline-specific, e.g. Dryad [9], Clinicaltrials.gov [11]; and (2) generalist repositories, e.g. Zenodo[1], OSF [2], FigShare[2]

- Genome Sequence Archive [7], *large-scale* repository handling big sequence data in a *high-throughput* environment; archives $\approx$ 16.9 PBytes since 2015
- Chinese National Gene Bank [3], archives $\approx$ 8.5 PBytes and provides public access ($\approx$ 0.13% thereof)

Resource requirements

Operating such an infrastructure for data activities in educational factories constitutes a huge burden for many organizations

[1] `https://zenodo.org/`, accessed 2022-10-24
[2] `https://figshare.com/`, accessed 2022-10-24

Related Work

- Spreadsheet applications [5] are available in many organizations, understood by many students, connected to a relational database

Machine-actionability

Addresses tasks that involve human actors and need for collaboration between those without input validation and storage strategies

- The OxCOVID19 Database [4] provides COVID-19 pandemic related data in a single relational database, aggregates openly available data

Open computation and -data

Reproducing queries at arbitrary times may be impossible since any modification on the dataset permanently replaces the old values, archiving subsets in generalist repositories to make them findable

Conclusion

We proposed a standardized data management method and data analytics process for educational factories:

- Specialized Data Repository
- Identified requirements for such a specialized Data Repository based on our discussions and interviews
- Prototype based on the existing DBRepo [8] approach
- Remaining challenges and impacts of the envisioned system (technical, societal)

Future work

- Address remaining socio-technical challenges (open vs. restricted data, graph-based data and triple-stores)
- Development of the auditability and transparency aspects of the Data Repository

Contact

Martin Weise
martin.weise@tuwien.ac.at

Data Science Research Unit
Technische Universität Wien
Austria

References I

[1] R. T. Fielding and R. N. Taylor. *Architectural Styles and the Design of Network-Based Software Architectures*. PhD thesis, University of California, 2000.

[2] E. D. Foster and A. Deardorff. Open Science Framework. *Journal of the Medical Library Association*, 105(2), 2017.

[3] X. Guo, F. Chen, F. Gao, et al. CNSA: a Data Repository for Archiving Omics Data. *Database*, 2020, 2020.

[4] A. Mahdi, P. Błaszczyk, P. Dłotko, et al. OxCOVID19 Database, a Multimodal Data Repository for better Understanding the Global Impact of COVID-19. *Scientific Reports*, 11(1):1–11, 2021.

[5] F. Nurdiantoro, Y. Asnar, and T. E. Widagdo. The Development of Data Collection Tool on Spreadsheet Format. In *2017 International Conference on Data and Software Engineering*, pages 1–6, 2017.

References II

[6] A. Rauber, B. Gößwein, C. M. Zwölf, et al. Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data. *Harvard Data Science Review*, 3(4), Oct 2021.

[7] Y. Wang, F. Song, J. Zhu, et al. Gsa: Genome sequence archive. *Genomics, Proteomics & Bioinformatics*, 15(1):14–18, 2017.

[8] M. Weise, M. Staudinger, C. Michlits, et al. Dbrepo: a Semantic Digital Repository for Relational Databases. *International Journal of Digital Curation*, 17(1):1–11, 2022.

[9] H. White, S. Carrier, A. Thompson, et al. The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment. In *Dublin Core Conference*, pages 157–162, 2008.

[10] M. D. Wilkinson, M. Dumontier, et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3(1), 2016.

References III

[11] D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide. The Clinicaltrials.gov Results Database - Update and Key Issues. *New England Journal of Medicine*, 364(9):852–860, 2011.