



Artificial Intelligence supported Tool Chain in Manufacturing Engineering

ITEA 3 – 19027

Work package 2

Extraction of knowledge and process information from
experience and data

Deliverable 2.1

Classification of relevant data and sources for production
systems

Document type	: Deliverable
Document version	: 1
Document submission date	: 2022-02-28
Classification	: public
Contract Start Date	: 2021-01-01
Contract End Date	: 2024-02-29





Artificial Intelligence supported
Tool Chain in Manufacturing Engineering
Project Coordinator: Johan Vallhagen, Volvo



Final approval	Name	Partner
Review Task Level	Dorchain, Marc	Software AG
Review WP Level	Deniz Bilgili	Ford Otosan
Review Board Level	Matthias Riedl	IFAK

Executive Summary

This document summarizes data sources found in production environment that are in close relation to the use cases brought by industrial project partners. The range of production data sources is very broad and specific to production unit and process. However, data sources can be classified to group data by type, origin, generation speed and format. The most common formats have been identified in this document and the most common data sources both in the online phase of the production as well as in the planning/improvement stages. Two main categories of data have been identified: human generated data and machine generated data. While in both realms data is finally generated by some sort of computer system, the key difference lies in format and predictability of data generation intervals. Human generated data is hard to predict, it covers design plans, production line layout, assembly structures and sequences. This type of data is generated as needed and updated sporadically as compared to automatically generated data. It is also characterized by large variability in the structure, as same description of an object can be done in multiple ways – for example 3D or 2D drawing of a part, textual description or visual marking of features and process parameters, structuring of assemblies can relate to function, division to subsystems, or represent arbitrary grouping that serves only part alignment with respect to each other. On contrary machine-generated data is usually of predictable format, and forms often data streams or repeatable structure but variable content. This allows not only expecting data flow related to well-defined timing, or event based, but also expressed in predetermined format and communication protocol. Which makes such data straightforward acquisition, storage and cataloguing possible.

In this document we have also identified formats that will be used for data representation in AIToC framework serving as standardized exchange format. This also lead to development of preliminary framework architecture, that would allow collection, storage, cataloguing and retrieval of data from a central repository. This architecture is designed to be modular, in order to allow data processing units to be independent and easily extendable.

Content

Executive Summary.....	3
1 Introduction.....	5
1.1 Data generated by humans.....	5
1.2 Data generated by machines.....	10
1.3 Current data use from production (Daimler, Volvo, Ford – bullet points or short paragraphs).....	10
1.4 AITOC project data usage vision (Daimler, Volvo, Ford – bullet points or short paragraphs).....	10
2 Assumptions about data sources.....	11
3 Classification of data sources in production.....	14
3.1 Sensor generated data (real world data).....	14
3.1.1 Production.....	14
3.1.1.1 Automation related.....	14
3.1.1.2 Human related (Daimler, Volvo, Ford).....	16
3.1.1.3 Statistical data.....	19
3.1.2 Supply chain and logistics.....	19
3.2 Human generated data (digital world data).....	20
3.2.1 Product and Resource Data.....	20
3.2.2 Process data.....	20
3.2.3 Layout data.....	20
3.2.4 Factory Data /Building Information Modeling.....	20
4 Data relevance evaluation.....	21
5 Common data features.....	22
6 AITOC system architecture proposal.....	23
7 Summary and conclusions.....	24
8 References.....	24

1 Introduction

1.1 Data generated by humans

During product development process, product plans are created. Those plans contain the geometric information (CAD data), as well as functional requirements, quality related specification, can include software component for products containing electronics, and finally testing requirements and procedure for assessing part's fitness and correctness of the production process. All this data is mostly created by engineers and experts involved in the product development cycle. The aim of the AIToC project is to support generation of this data using expert systems, and data processing and reasoning modules. Data generated by humans is quite diverse as it spans from well-structured and computer processing oriented CAD data, to specifications that can be prepared in natural language in human friendly but pdf, word, power point formats, but at the same time those formats are very complex to process by automated systems, as they lack ontology and well defined structure and syntax.

Among industrial partners involved in the AIToC project the following data formats for CAD data has been identified:

- Catia files
- Siemens NX/Unigraphics
- JT
- Collada
- Creo

In addition common CAD formats utilized in industry are:

- Solid works
- Autodesk Inventor
- AutoCAD
- Revit
- Pro/E

The exchange of data between proprietary software is not straight forward and data is often lost using exchange formats. Among the available CAD data exchange formats one can find:

- STP
- IGES
- Parasolid
- VRML
- DXF
- STL
- 3D PDF
- JT
- Collada
- glTF

Selection of one of the above mentioned exchange formats as a standard CAD data carrier for the AIToC framework is important step, as it influences all further CAD data processing capabilities of the AIToC tools and modules. It is desired to identify preferably one common format that would allow collecting geometrical data about assemblies, individual parts, material information, geometrical constraints, and finally labels and annotations allowing to give ontology connections between geometrical features and processes required to build them or utilize them in the assembly process. Table below summarizes pros and cons of the common exchange formats.

STEP	
STEP is popular neutral CAD format as it was standardized by ISO committee in 1994 under the official name "ISO 10303-21", and was designed from the beginning to be the international standard. STEP stands for S Tandard for the E xchange of P roduct model data.	
<p><i>Pros</i></p> <ul style="list-style-type: none"> • Stores data using a mathematical representation of curves (referred to as NURBS) to give a perfect surface definition • Most widely used and accepted neutral format today (standard across many industries) • Developed by ISO • Good file compression (making it more ideal for sharing over the internet) • No loss of quality with the source files in terms of details • Allows downward compatibility (Example: A file created in Autodesk Inventor 2016 can still be used in Autodesk Inventor 2015) 	<p><i>Cons</i></p> <ul style="list-style-type: none"> • No materials or textures as STEP focuses on shape and form • Slow to release new updates quickly • No lighting or camera data • Can't be directly rendered as it needs to be processed by a software algorithm that converts the mathematical surface representation into a series of triangles • Cannot contain parametric intelligence and feature history
STL	
STL, which stands for Stereolithography, was designed back in the 80's when file size was a BIG deal. Similarly, it has only been within the last decade that the "model is master" has been adopted by the industry. The STL file was built in a time where things like texture, small details, and EVERYTHING you ever wanted to know about a part was captured in the drawing, not the CAD file. The CAD file, up until recently, was supplementary. But now, you can store measurements, material information, texture, and loads of other information about the part.	
So why do we use stuff like STLs today? Well, it's easy. The code is written. It's inexpensive to continue to use something that's been around for so long. And it still has its value today. Again, great for rapid prototyping and small file size.	

Pros

- Great for prototyping or gathering estimated volumes and measurements from
- Nearly universal and very commonly used
- File size

Cons

- Not something you'll want to use if you are going to need super-precise measurements. It is not a precise representation of a part
- Geometry resolution can cause issues in manufacturing (a circle will never be a true circle)
- Only describes the surface geometry of a 3D object
- Does not represent Color, Texture, or other common CAD Attributes

IGES

IGES, which stands for Initial Graphics Exchange Specification, was the first neutral CAD file format invented towards the late 1970s, early 1980s. This format is mainly used for surface geometry and design work. While IGES is widely supported, it has a hard time computing complex parts' faces, so it will guess and break models into surfaces, resulting in the user having to repair to get to solid body in some instances. Its limitations and the variety of better alternative neutral formats are resulting in engineers moving away from this file type.

Pros

- Widely supported

Cons

- Files are imported as solid models, not surface geometry
- Often gets translated with gaps between the surfaces, missing faces, and occasionally surfaces with faces in the wrong orientation
- Cannot carry MBD (Model-Based Definition) data, nor can it carry PMI (Product and Manufacturing Information) data
- IGES files often have to be repaired resulting in engineers having to spend several hours a week fixing design data

3D PDF

3D PDF format is considered as more of a universal choice for sharing and exchanging CAD prints as anyone with Adobe Reader can open and view it.

<p><i>Pros</i></p> <ul style="list-style-type: none"> • Easily viewable on computers, smartphones, and tablets • Great for sharing with those who don't have CAD software or viewers, as PDFs only require Adobe Reader to view files • Can be used to help reduce file size in order to send via email 	<p><i>Cons</i></p> <ul style="list-style-type: none"> • Slow performance • Very limited capabilities
<p>Parasolid</p>	
<p>Parasolid is a geometric modeling kernel that can be licensed by other companies for use in their CAD software. This format's capabilities include model creation and editing utilities, advanced surfacing, thickening & hollowing, blending & filleting, and sheet modeling.</p>	
<p><i>Pros</i></p> <ul style="list-style-type: none"> • Great CAD export option for engineers using SolidWorks or NX 	<p><i>Cons</i></p> <ul style="list-style-type: none"> • Not a standard format • Cannot communicate and migrate 2D data such as lines and arcs • Has to be licensed
<p>VRML</p>	
<p>VRML, which stands for Virtual Reality Modeling Language, is a standard format used to represent 3D interactive vector graphics. VRML files are in plain text and are useful for transferring over the internet more quickly. This format uses a polygonal mesh to encode surface geometry & can store appearance-related information (such as color and texture).</p>	
<p><i>Pros</i></p> <ul style="list-style-type: none"> • Compress well using gzip, making it more useful for transferring via the internet 	<p><i>Cons</i></p> <ul style="list-style-type: none"> • Has not received wide acceptance
<p>Collada</p>	
<p>Collada is a 3D file format used more heavily within the video game and film industry. This format supports geometry, appearance-related properties, materials, textures, and animations. In 2013, it was adopted by ISO as a publicly available specification, ISO/PAS 17506, which resulted in lots of 3D modeling software supporting the format</p>	
<p><i>Pros</i></p> <ul style="list-style-type: none"> • Supports kinematics and physics • Supported by lots of 3D modeling software 	<p><i>Cons</i></p> <ul style="list-style-type: none"> • Hasn't stayed up to date, resulting in some shifting more towards other formats
<p>DXF</p>	
<p>DXF, which stands for Drawing eXchange Format, is the neutral format from AutoCAD that can cross platforms (unlike the DWG format). DXF is a vector file that stores 2D drawings, meaning that you can edit individual elements that make up an image. DXF files are widely used as DWG since it is supported by most CAD programs.</p>	

<p><i>Pros</i></p> <ul style="list-style-type: none"> • Open-sourced and almost every CAD software supports it 	<p><i>Cons</i></p> <ul style="list-style-type: none"> • DXF files are usually larger in size • Only retains information such as line work, dimensions, and text • Does not support application-specific information
<p>glTF</p>	
<p><i>Pros</i></p> <ul style="list-style-type: none"> • Open-sourced and support for it is growing • Small file size as it is optimized for Internet and network use • Can store animations, lights, assemblies, materials, mesh transforms • It is extendable, so adding meta information is straightforward • Based on JSON for structure information carrying and simple binary format for the numerical data • Supported by web standards and implemented in various 3D web viewers • Portable as it can include all information in a single file 	<p><i>Cons</i></p> <ul style="list-style-type: none"> • Design tree information is not natively saved, it would need to be added as extension if it is required • Geometry is stored as meshes, therefore feature information like round holes, flat sections, arcs need to be either saved as separate set of metadata or need to be extracted from mesh data using relatively complex process.

After selecting the standard CAD format for the AIToC framework it is important to define tools and processes that would allow exporting data from native formats utilized by the industrial partners to the common format and vice versa. In the proposed framework architecture this task would be dedicated to connector component.

In addition to the geometrical information, which also contain hierarchical assembly structure, eBOM and mBOM are commonly utilized to define process centred assembly structure. Bill of material formats can be based on Microsoft Excel, csv, or xml formats. Their structure differs between industrial partners and therefore those data files require mapping to standard input/output format that will be utilized in the AIToC framework. Mapping of the various resource files should be performed with a dedicated tool, that would ensure data integrity and interlinking with other data, for example coming from PLM/PDM systems or CAD software. Referrals to standardized parts as well as to internally produced parts can be done using part numbers. It is however important to maintain unique part numbering system that would allow problem free identification of concrete part and its revision solely by the part number.

1.2 Data generated by machines

Machine generated data refers to data that is created automatically in some process. Machine generated data can be considered as outputs from various robots, sensors, instrumented tools and even camera systems, where data generation is based either on events or fixed/variable time step that is predictable and repeatable. Machine generated data is often a stream that can contain endless production data or chunks representing certain time span.

1.3 Current data use from production

The considered use case of Daimler Buses is a station of the final assembly line of coaches in the plant in Neu-Ulm. Due to a high variety and a high cycles time most of the work is done manually. Currently, no data of the manual processes are collected directly, only quality issues are documented manually in the enterprise planning tool.

Ford's production line of the brake disc is provided as a use case and is under construction. Manufacturing data will be available end of Q1 2022. Currently, laboratory test data are available for drilling operation. Variety of operations will be increased with time. The data can be shared anonymously. Error codes from active production lines, motor currents/voltages of some robot arms and specific vehicle IDs can be collected in the current operations. There is no systematic data collection in manual processes at Ford.

At Volvo currently production related data is mostly processed manually. Geometric measurement data like point clouds, or laser scans are collected to prepare new production line layouts. Such data collection is manually operated but data are collected automatically during the process. To use the data for representation of the production environment, some amount of manual data manipulation is needed. Nevertheless, this effort allows accurate representation of space to be included in planning operations.

In all cases during the production PLC-driven robots, conveyers, etc. use signals to communicate with each other to maintain production line synchronization and flow of products. While robots and conveyers are programmed manually, this operation is done once and then during operation, such devices provide data automatically and this data is available for other devices. In transport, variation handling, and quality check domains, there are many semiautomatic systems for data collection and processing. Mostly focusing on tracking location of particular parts or subassemblies. Product issues are collected manually in free-text form, therefore such issue logs are not well suited for automatic processing. Nut runners and other torque type of wrenches automatically store applied torques to screws/nuts and those values are used for quality check and documentation.

1.4 AIToC project data usage vision

Based on the considered use case it is intended to collect some data of the manual process automatically. Therefore, this pilot station will be installed with additional sensors (e.g., AR glasses, touch screen or button, camera (on/off), QR Codes/RFID, data gloves) to identify:

- Part position
- Tool position

- Carrier and shelves position
- Ongoing task

After the brake system production line is constructed, these data will be collected in real-time:

- Electrical current from each electrical motor
- Acceleration data
- Spindle tightening torque
- Oil temperature and pressure
- Ambient temperature
- Air flow and air pressure of pneumatic systems
- Tool number used at work
- Number of blades of the tool used at work
- Total number of machining processes with machining time of each tool
- Percentage of tool life by total number of machining cycles
- Selected machining program defined according to the part recognition algorithm
- Indicating the current operation (clamping, loading, handling, unloading, transfer)
- Machine generated error codes

In the course of the AIToC project the use of data that is currently generated in production will increase. First of all, point clouds and laser scans should be directly usable in WP5 for layout planning minimizing manual data preprocessing, or ultimately eliminating such need. Data from robotic stations should be used in FMU creation from measurement data, so that ready FMUs can be used in new production planning processes. Assembly instructions, that are currently handled on paper or using part lists, should become created automatically from digital data and presented in digital form, following changes to the product and considering variant handling. This also implies use of worker assistance based on the automatically generated data.

2 Assumptions about data sources

Data source in production according to the AIToC are defined as: Sensor generated data (real world data), human generated data (digital world data), and simulation data (virtual cyber world). This corresponds to the definitions made as part of the WP3¹ in the AIToC project.

¹ see deliverable “D3.1 Context Definition_Final”



1. Categories of data sources

Real-world (e.g. a factory plant) mainly means sources of data streams from sensors or other IoT devices where Digital World mainly means digital artifacts that created by humans in large number of work flows. Key difference between real-world and digital world lies in the origin of the data – real-world data is produced automatically by devices and computers and the data production rate and time is based either on events or time constraints and intervals. Therefore, this type of data is forming a stream that will represent relatively simple data structure of repeating data blocks. It also contains measurements that represent physical qualities or software states. On contrary the digital world data is produced in far less predictable way. Design plans, requirements list, production plant layouts, CAD files are good examples of digital world data, which is mostly prepared by humans. This type of data contains representatives like CAD data or part lists that are already presented in one of the standard exchange formats allowing easy processing and parsing. Nevertheless, human factor included in those data sets makes them more difficult to process as compared to real-world data. For example, design tree of a CAD file is very specific to workflow of a specific engineer. One can first extrude main body parts and then plan holes as individual cut out features, another can integrate both main body and holes as single CAD operation still producing exactly same final part geometry. Rounded or chamfered corners can be result of separate edge rounding/chamfering operation, or can be result of extrusion of a sketch that already contains them. This means that feature recognition of CAD data and digital world data in general needs to take into account variability of the workflows and all possibilities of creating certain features using different methods.

Another specific to the digital world data example are descriptive documents, that are stored as texts, figures, tables, which are human readable, but are cumbersome to process automatically as they do not contain proper context definition and markup. This implies that data pre-

processing of the digital world data is concentrated on formatting and labelling of data elements to add semantic meaning to the data that later on can be mapped to internal AIToC data format. The virtual cyber world is also possibly relevant for the project but will be currently out-of-scope for this document. It is just important to note, that in the simulation domain, there is large number of commercial software specialized at different simulation tasks, and producing data in often proprietary formats, that are difficult to decipher. Nevertheless, such software can also produce digital data outputs that are in some sort of text format or open binary format, that can be processed automatically. Such data would be usable by the framework, however it raises the question if source simulation files which cannot be understood by the framework should be also stored in AIToC data warehouse for maintaining completeness of the company data and version tracking or should it be only referenced and stored on another platform.

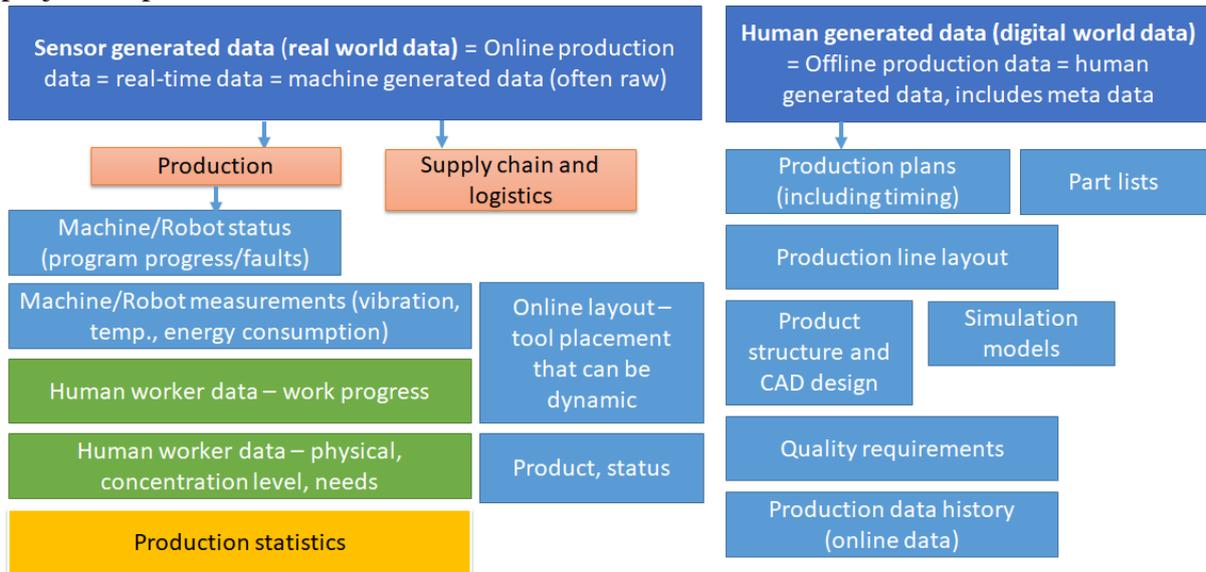
All data types that are considered relevant to the AIToC framework should be represented by internal AIToC format and should be available in the internal AIToC data exchange format for all the framework components. This leads to the architecture design presented in one of the following sections. Data homogenization and translation to internal data format is one of the important data pre-processing tasks that need to be one of the first tasks in the implementation stage of the project.

The preliminary assumptions about data sources can be summarized as follows:

1. Data source is consistent and can be read completely without input output errors.
2. Data can be incomplete, contain empty fields or null values.
3. Data must be accompanied by meta information. In case the data source directly does not provide meta information, the user interface for creating data connection should enforce user to provide sufficient meta information.
4. Data low-level format is well-defined.
5. Time representation is defined with the time zone or if the time zone is missing it is in the UTC format.
6. Real-world data is expected to be mostly numerical values, text values should be enumerable and fully defined in data connector.
7. Text data elements that are not defined can be substituted by predefined replacement value (for example null, predefined string, or predefined number).
8. Data is primarily available in semi-structured text format – plain text, csv, xml, json, etc.
9. Text data encoding is always known if it is relevant.
10. In case of binary data, the data structure is well defined and known.
11. Data can be in a form of a file or TCP/IP stream.
12. In case of data file the timestamps from file system are ignored as unreliable.
13. Data source devices are assumed to have time synchronized and timestamps inside files or data is correct.

3 Classification of data sources in production

The classification of data source in AIToC is following the illustrated separation of the two main types of data sources – real world and digital world. The following image illustrates the structure and main kind of data sources that will be handled or will be available within the project scope.



2. Data sources in production

3.1 Sensor generated data (real world data)

3.1.1 Production

Machines or machine tools including robots do produce status data by default even if not all of them are cyber-physical systems (CPS). The status might include basic information like power on/off, working hours, location etc.

3.1.1.1 Automation related

Machines or machine tools including robots may be equipped with certain sensors that are able to deliver – in most cases continuously – data that is created during the usage (runtime) of the machine. Typical examples for this kind of data are vibration or rotation (e.g. at drilling machines for electro spindles), power consumption, temperature of the environment or specific parts of the machine like the board or the motor.

Specific examples of data produced by specific device related to industrial partners of AIToC are listed in the table below:

Table 1 Typical automation data collected in industrial process

Device	Input data	Output data	Device data/comments
Portal robot	Position, movement instructions and	Safety signals (logical and	CAD model of the robot,

	<p>program, laser gates (stops the process if someone is too close)</p>	<p>physical), control signals, inductive sensors for tool and buffer detection</p>	<p>adaptive for specific use case, (kitting, assembly operation),</p>
<p>PLC-controlled equipment</p> 	<p>Assembly process requirements from documentation and quality description (angle, tightening torque, force)</p>	<p>digital signal output (angle, tightening torque, force)</p>	<p>Nut-runner</p>
<p>Conveyor</p> 	<p>Design speed, start and stop positions, product position on the conveyer, emergency stop button</p>		<p>Control data is set once and does not change during the production so it is not available as output.</p>
<p>Collaborative Robot</p> 	<p>Program, cameras input (images), training images, instructions for robot operation</p>	<p>Detected object type, detected object location and orientation, human position and status, movement, status</p>	<p>Used in lab environment and currently main use case is for kitting</p>
<p>Automatic quality check</p> 	<p>Training data set (images with labels)</p>	<p>Digital signal output. Cameras and sensors detecting missing parts or misplaced parts, checkpoints for specific production steps, quality campaigns</p>	

Daimler requires that all new machines are able to provide data externally through OPC UA[1] protocol. Volvo supports OPC UA standard as well. That gives access to all sensors from the machine (if the machine vendors support the access to all data). TWT uses MODBUS/TCP [2] for communication between PLC and external elements (sensors/actuators). In addition commonly used is MQTT [3] protocol that is light-weight and specially designed for IoT.

Moreover, it can also be used for the data collection from production equipment. PROFINET is another protocol found in AIToC use cases. In the AIToC project the base protocol will be MQTT, and the secondary protocol that will be supported will be OPC UA.

Complex Event Processing (CEP) is a technology that might be interesting to use in conjunction with the mentioned protocols because it is focused on direct real-time use of the data stream on the protocol level without the need to first store data in any kind of database. Examples of tools to execute CEP are Esper (<https://www.espertech.com/esper>) or APAMA (<https://www.apamacommunity.com>) from Software AG where a free community license is available. Such platform might be useful in AIToC implementation considering processing of the data that is produced by IoT and production equipment.

3.1.1.2 Human related

Ford uses fully automated lines with no human workers for their use case. Human related motions are not recorded by camera or other sensors on a regular base due to restrictions of the people. In general, storing personalized data of the operators are usually not allowed in factory environments. For specific projects and for a limited time recording of data is possible, but it has to be confirmed by involved operators and the union.

Criteria like physical condition, concentration level could be monitored on a volunteered base, e.g. by pushing a button or using some simple user interface to give worker option to mark his attitude. Manual assembly processes are difficult to monitor automatically in a linear fashion. Currently, there are checkpoints (“quality gates”) on the assembly line. Each time a truck or bus passes such a checkpoint, it gets checked if parts are assembled correctly. These checkpoints are used to monitor the quality of the truck or bus and can be understood as some sort of progress monitoring. Otherwise, the assembly line is continuously moving, there are no physical buttons the operators need to press to move on.

In special cases, like security relevant screws and joints, the operators need to document the tightening process. It needs to be traceable, and therefore could be used to indirectly monitor manual process progress. Such tracking is done by the use of PLC-controlled tools. There are also buttons in use for operators to verify that the task has been completed.

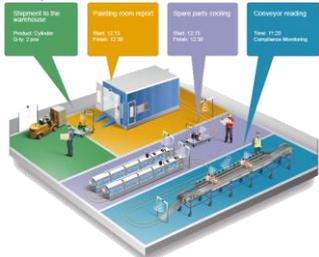
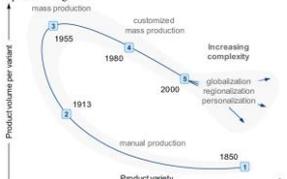
In case problems are identified in production process, there is a system at Volvo, called PIL (Product Issue Log). Daimler is using similar functionalities integrated in the enterprise resource planning tool SAP. Production engineers can report problems in production in such a system. Production and product engineering department can then work together on solving the problem. Problem in one factory is then analysed also in other factories to verify if it is local or systematic problem.

Problems and their number as well as specific errors in production are monitored. Depending on impact of the problem and the quantity, there are different consequences and actions for production. But anyway the final decision on actions needed is based on human.

No specific methods to monitor the operators are in use. There are however team leaders in production, who have responsibility for the operators in their team.

AIToC industrial partners have summarized in Table 2 typical data related to human that is collected in production process.

Table 2 Typical human related data collected in industrial process

Type	Input data	Output data
Layout 	Production specification and requirements	Point clouds, laser scans, points of use – input output points where material is delivered or picked up, layout
Product issue Log 		Deviation card, quality issues, audit (most cases free text description), statistical data
Manual assembly station 	Assembly instructions, part numbers, time allocated for each task, quality requirements	Measurements from instrumented tools, manual reports on needed rework activities
HMI (Human Machine Interface) 	Assembly instruction data, 2D/3D CAD - images/models, production planning data, assembly sequence, part lists, mbom (manufacturing bill of materials)	Information is only provided to the user but no data is collected back.
Variation handling 	Variant description, every order is considered separate variant for Daimler	Statistical data, production control, logistics, rework activities (free text information) – paper and electronic versions are kept

Quality check 	Quality criteria specification	Detect missing parts or misplaced parts, checkpoints for specific quality campaigns (no digital data)
Tooling and equipment 	Product specification, product and assembly CAD data, process instructions and specification	CAD data (no operational use), for old equipment no models are available Mismatches can happen between CAD and real tool if some adjustments are made during the process.

Human related data can bring information about production progress reporting, but also human well-being and location within factory environment. Nevertheless, privacy issues often arise while discussing human-centred data. GDPR requires consent of the human for automated data collection as well as clear specification of what kind of data is collected and how it is used. In addition, more demanding regulations are imposed by unions and work contracts. Therefore, data about manual operations is rather sparse and limited to sensors placed just before or after manual assembly work cells. This data is available from sensors and cameras monitoring products and parts coming in and leaving assembly cells. From ergonomic evaluation perspective, information about tool travel during manual operations as well as human body movement would be of interest. Such data cannot be currently collected constantly during production, but can be collected in specially organized sessions once in a while shedding some light on real-life ergonomics and not only design phase assumptions about how tasks will be carried out.

Well-being information focused on monitoring stress level, concentration, muscle fatigue or estimation of joint loading could be used to design more optimized work cells. It raises however privacy concerns, as such monitoring would have to be carried out using sensors attached directly to human body (pulse measurement, body temperature, etc.) and cameras combined with machine vision systems for analysing concentration level based on eye movement, face expressions; even keystroke dynamics can be used to estimate agitation level of a computer user [4]. This leaves current state of the human parameters estimation in a level, when data can be collected only during organized trials with users consent, and online through indirect measurements that are carried out on products and machines rather than on humans, but can be analysed and interpreted to provide information about workers. For example number of faulty components produced, amount of operations done per time unit, delays in reactions to instructions, etc.

3.1.1.3 Statistical data

This includes statistical data that is assembled during the production lifetime, e.g. the number of parts being produced or assembled during a specific production period. Production planning system is available in each manufacturing company (SAP for Daimler, Volvo uses internal system), all the major production events are recorded in the system. Those systems collect quality information. Despite big datasets are available, it is not straightforward to retrieve the data.

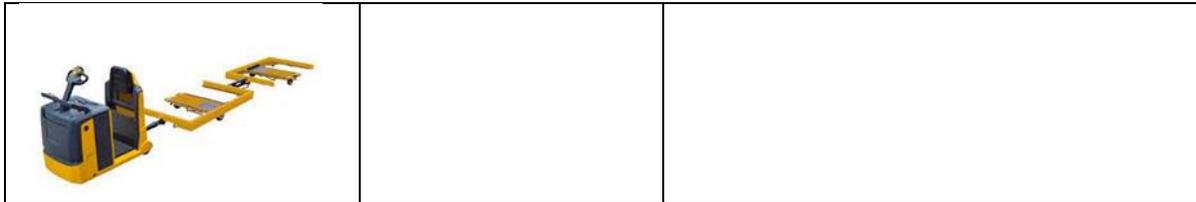
3.1.2 Supply chain and logistics

What transport systems are used (autonomous, manual, human operated, remotely operated), do they collect any data currently, if yes, what kind.

How demand for transport operations is provided, is there some system for this or is it manual communication

Table 3 Typical supply chain data collected in production process

Device	Input data	Output data
Transport Robot (AGV) 	Transport source and destination points, transport schedule, dynamic transport requests	Transportation data (location, obstacles, destination, laser map), difference between line follower and dynamic path planning (route), odometry (positioning), 3D-camera (obstacles detection)
Inventory 	Refill plan, kitting map	Inventory status, localization, kitting data
Pallets 	Transport source and destination points, transport schedule	Inventory status, localization, kitting data
Tugger Train / Human driven delivery	Demand, source and destination points, schedule, dynamic demand for products (one cycle time in advance)	Checkpoints marking delivery information Whenever there is delay (not in time) event is raised. Connected to production planning system.



3.2 Human generated data (digital world data)

Human generated data is basis for the current factory operations. Designs, specifications, layouts are all produced in a human-centered process. Product and Resource Data. No real-time data is included in this data sets. Most data is generated and used sporadically.

3.2.1 Product and Resource Data

Product and resource models are generated by CAD and stored in a data base or on a project drive. In the introduction key digital formats for CAD data have been described as one of the core engineering outputs of humans in production planning. Additionally for every product a detailed list of parts that need to be assembled is provided in a form of eBOM or mBOM list that tend to be standardized within organizations, but not world-wide.

3.2.2 Process data

All automated or manual process are usually described digitally and stored in a database or on a project drive. This data is company specific and lacks standardization. In some case Automation ML is used, Automation ML is also utilized as exchange format. JSON and proprietary formats are used for transmitting the data. Binary data is used when data transmission rate is critical (increased efficiency in numerical data transfer). Real-time data is not well suited for Automation ML or other XML-based formats due to its size overhead.

3.2.3 Layout data

Layouts are partly generated by BIM (Building Information Modeling) and partly by CAD. Dominating format of the data is 2D layouts. Factory environment is most of the times reused for new production processes. 3D layout of the factory can be provided as a plan. But building plans are only in 2D. Layout can often be as raster picture (PNG, JPEG).

3.2.4 Factory Data / Building Information Modeling

Building information modeling (BIM) is a process supported by various tools, technologies and contracts involving the generation and management of digital representations of physical and functional characteristics of places. Building information models (BIMs) are computer files (often but not always in proprietary formats and containing proprietary data) which can be extracted, exchanged or networked to support decision-making regarding a built asset. (source: https://de.wikipedia.org/wiki/Building_Information_Modeling)

Micro station software is used for factory planning. AutoCAD is another popular software used for factory planning.

4 Data relevance evaluation

Table 4 shows what data sources are available and marking relevance of the data sources to AIToC in three level scale (Unimportant, Important (support data), Essential (key importance) for the WP evaluation. As summary general importance introduces additionally Very important to express rating between Important and Essential, as well as Useful for expressing intermediate level between unimportant and important ratings.

Table 4 Data sources importance for WPs and generally for AIToC

Data source	Importance for WP 2	Importance for WP 3	Importance for WP 4	Importance for WP 5	General importance
Sensor generated data – real world data					
Portal robots	Unimportant	Unimportant	Essential	Essential	Important
PLC-controlled equipment	Essential	Unimportant	Essential	Essential	Essential
Conveyor	Unimportant	Unimportant	Essential	Essential	Important
Collaborative robot	Unimportant	Important	Essential	Essential	Essential
Automatic quality control	Unimportant	Unimportant	Essential	Essential	Important
<i>Human monitoring</i>	Unimportant	Important	Unimportant	Important	Useful
<i>Tool position monitoring</i>	Essential	Important	Unimportant	Important	Important
Human related data					
Factory layout	Unimportant	Essential	Important	Essential	Very important
Product issue log	Unimportant	Essential	Important	Unimportant	Useful
Manual assembly station	Unimportant	Essential	Unimportant	Essential	Important
HMI	Unimportant	Essential	Unimportant	Unimportant	Unimportant
Variant handling	Unimportant	Essential	Important	Important	Important
Manual quality check	Unimportant	Essential	Important	Important	Important
Tooling and equipment	Important	Essential	Important	Important	Very important
Supply chain and logistics					
AGV	Unimportant	Unimportant	Important	Essential	Useful
Inventory	Unimportant	Unimportant	Unimportant	Essential	Unimportant

Pallets	Unimportant	Important	Unimportant	Important	Useful
Tugger train/human driven delivery	Unimportant	Unimportant	Unimportant	Essential	Unimportant

As a summary PLC-controlled equipment data as well as data from collaborative robots are the essential for the AIToC platform to comprehend. Factory layout, tooling, and automatic quality control are second most important data sources for AIToC. Based on current use cases portal robots, conveyor systems, tool position monitoring systems, manual assembly stations, variant control, and manual quality control are important data sources for industrial processes. The remaining data sources identified in Table 4 are either useful in some particular cases or can be completely ignored in current focus of AIToC. This data importance summary allows to prioritize implementation of connector modules that will interface AIToC platform to data sources allowing straightforward data acquisition and preconditioning.

5 Common data features

Based on the previous data sources examples, a few common data types can be identified. First of all real-world data aka sensor generated data comprises of tabular entries. Each new data row contains same amount of columns that represent same data type for each respective column. This type of data is well suited for simple table storage, that is offered by any type of database engine. This type of data can contain multiple rows with exactly same values, or data columns might show specific patterns, that can be linked with the machine function, process parameters and specific production events. This type of data can be either produced constantly with fixed time step, or produced based on events in unscheduled intervals. In the first case time series analysis methods are the most appropriate analysis tools. In the second case for event based measurements, event based processing would be more appropriate.

Human related data represents set that is more diverse and often contains unstructured data. This makes such information far harder to store and retrieve in logical order as well as such data processing is much more complex. Nevertheless, subtypes of human related data can be represented in tabular format similarly to real-world data. All types of logs, can be represented as table containing fixed set of fields, representing timestamp, location reference, process reference, machine reference, and the actual log data that often might be free text. From storage, retravel, and transfer perspective such data is no different to automation data. Reasoning and processing of such data is however special case that requires natural language processing module.

CAD data representing either layout, product assembly, or part designs can be stored in a common format. For such data to be meaning full, a set of data labels, meta information, and annotations needs to be provided. At this stage of the project, we assume that data labelling and marking is firstly done manually, but based on limited set of keywords and data formats, with time, data labelling will be upgraded to semi-automated system that reuses labelling templates

created earlier manually to discover features and link them with specific labels and markings. Finally, labelling will be done fully automatically with just sanity check made by human workers and occasional expansion of labelling dictionary to include new processes, features and subsystems that will emerge in the future. Annotations and markings of geometrical features can be part of the CAD data and can be stored together with the CAD data of particular part/product/assembly/layout, etc. Labelling that should bring semantic meaning to annotated geometrical features can also be included to the CAD formats considered by AIToC (glTF, COLLADA), however it might also be possible to separate such labelling information and include it in Automation ML format. The decision on this issue is still not clear as both options have some benefits and disadvantages.

Tooling and equipment fits to relational database data storage method.

6 AITOC system architecture proposal

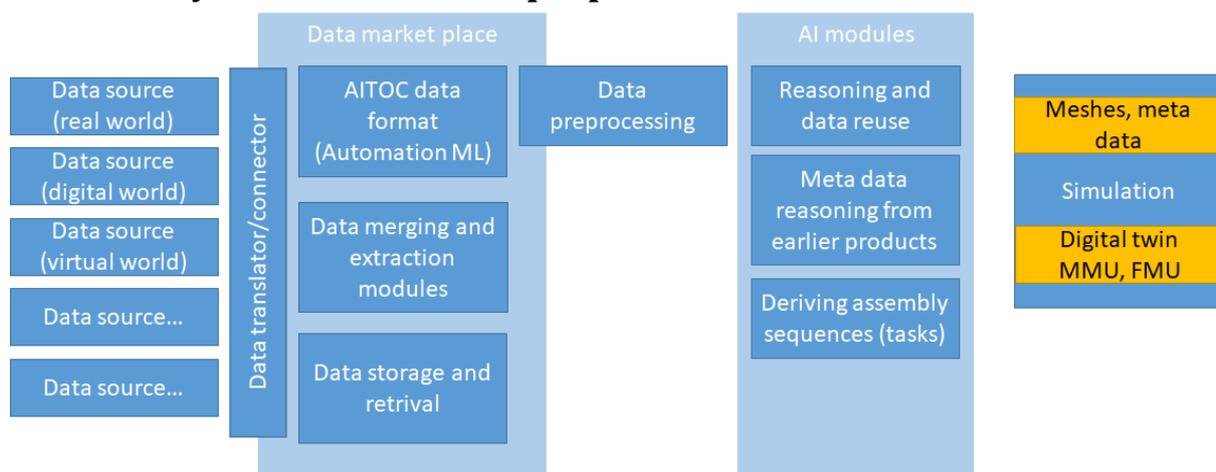


Figure 1 Components in AIToC

Figure 1 depicts proposed AIToC system architecture focusing on data sources and consumers. Among data sources there are three distinct worlds: real, virtual, and digital representing measurement data, simulation data, and human generated digital data respectively. Each type of data sources requires specific interface that leverages data stream size, real-time filtering and pre-processing, and security of data communication. While human generated data can represent large volume, it is not time critical input and can be delivered offline, just as it is produced. Support for common CAD format in this area is critical to limit compatibility issues of the framework components. As this is an initial architecture proposal, it is assumed, that the CAD data will be provided in one out of 2-3 selected file formats that should be most complete to represent this type of information. This imposes that file mangling is assumed on the CAD tools or external exporters and not being part of AITOC development process. AIToC will merely define formats that are standard for data exchange within framework.

Real-world data aka measurement data was summarized to be of similar format among different use cases and sources. This allows to select 2 main industrial data formats for transporting this data. Security of the data should be maintained through protocol level encryption, access

restrictions to the system, and isolation of the data acquisition system and processing system from external access whenever it is possible.

Virtual world data – simulation results and models – should always be stored in native source formats for maintaining full readability and compatibility with the software used to generate such data. Nevertheless, in parallel, data copy in Automation ML, or other open and clearly described format should be available for actual data processing and utilization by AIToC framework. This implies that there is at least one directional transformation from proprietary or close-sources simulation models format to standard exchange format.

Data market place is the center element of the AIToC architecture. The storage mechanism is not relevant for the AIToC consumers and providers, but the data exchange format and protocol need to be defined. The storage of the data will be done in three levels, depending on the data type. Annotations, meta information, ownership information, and interconnections between data and their sources will be stored in relational database. CAD files, and large complex data elements will be stored in key-value blob storage mechanisms, which are robust and provide good indexing and data retrieval performance, and they offer scalability that is required for constantly growing data sets. Finally, structured, time series data can be either stored in relational database tables, dedicated time series databases or as key-value pair sets, depending on the processing needs for the data and depending on the use case needs. If fragments of data are often required and not complete data sets, then relational database or time series database seem to be a better choice, if large complete sets are required for analysis, then key-value storage or file based storage tend to be more performant and scales better. Analysis will be concluded to determine optimal storage mechanism for the AIToC data.

AI modules will utilize unified access mechanism that will contain protocol and data format definition that should suit all possible use cases. This protocol will allow both data retrieval as well as saving new data sets maintaining relationship to the source data.

7 Summary and conclusions

Classification of common data sources in production lead to definition of three realms where data is created: virtual, digital, and real. Data types have been identified that match document type of data, witch needs to be always processed as one set, and fragments of such data is not sufficient for analysis; time series based data, that has fixed structure and data types, but is not self-contained, and can be analysed based on time window of interest or based on data patterns that need to be identified. Finally, annotation and relation data between data sets and company production equipment, tools, products, that exhibit relatively low volume of data but high number of references to other elements to create dependencies between products, parts, actions, etc.

8 References

[1] „OPC UA Online Reference”, 27 styczeń 2022. <https://reference.opcfoundation.org/>



Artificial Intelligence supported
Tool Chain in Manufacturing Engineering
Project Coordinator: Johan Vallhagen, Volvo



- [2] „MODBUS protocol specification”, 27 styczeń 2022. <https://www.modbus.org/specs.php>
- [3] „MQTT specification”, 27 styczeń 2022. <https://mqtt.org/mqtt-specification/>
- [4] R. Solanki i P. Shukla, „Estimation of the User’s Emotional State by Keystroke Dynamics”, *Int. J. Comput. Appl.*, t. 94, nr 13, 2014.