MaiML
(Measurement Analysis Instrument Markup Language)

Proposal for a comprehensive file format
for the measurement and analysis of Surface Chemical Analysis data

Backgrounds

- Needs for comprehensive representation of diverse, continuously developing measurements and analytical techniques in surface chemical analysis (SCA), in which various beams of electrons, ions, neutral atoms or molecules, or photons are used on the specimen material over the surface, and surface-related signals.
  - Reference: formats for individual measurements/analyses

    Analysis Instrument Markup Language (AniML): Standard Language for analytical chemistry and biological data, etc.
- Surface and surface-related Measurement analysis in materials development, including catalyst, requires comprehensive handling of multiple processes
  - Production of raw materials and materials

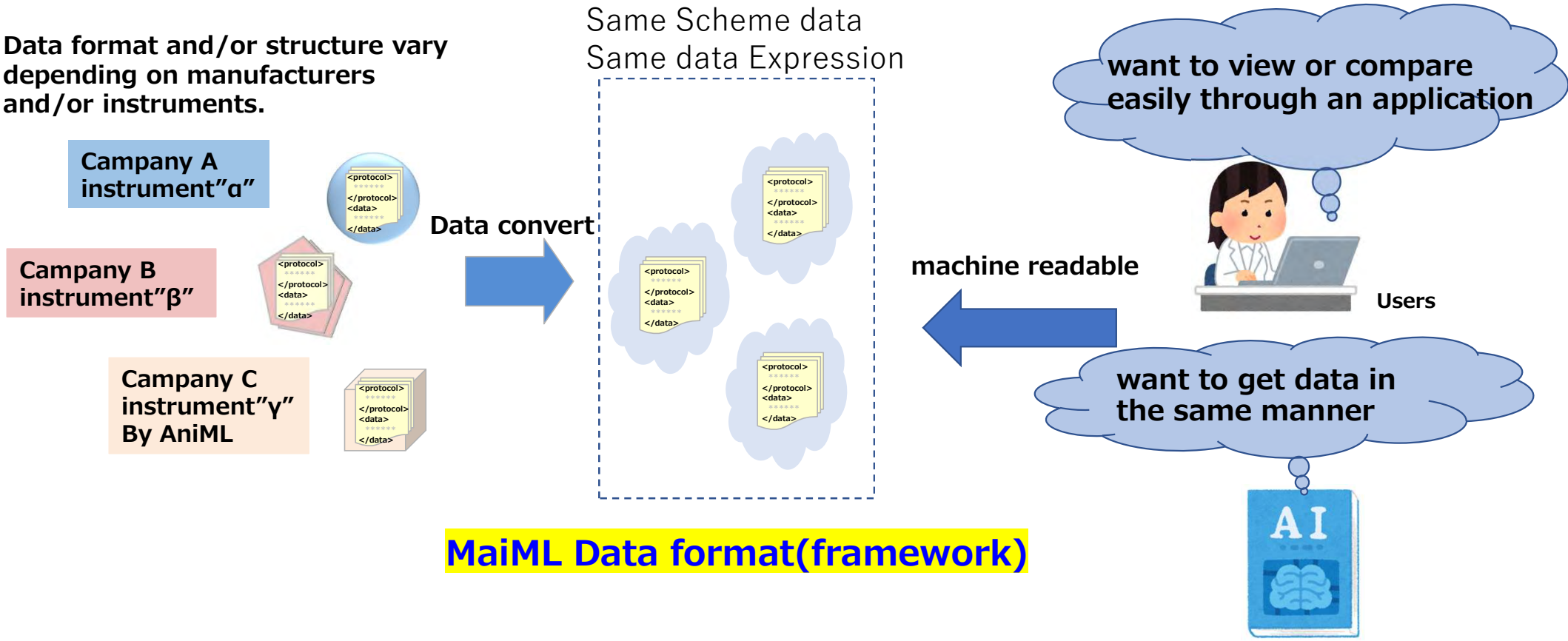
  → Preparation of samples for (multiple) metrological analyses

  → Results of (multiple) metrological analyses

  → (integrated) post-processing and analysis of (several) measurement analyses
- Data uniqueness must be guaranteed, and falsification must also be detected.
- Need for both data integrity and concealment (corresponding to open/closed strategies)
- Need to accumulate and integrate various data in SCA to enable big data analysis, such as AI and data science.

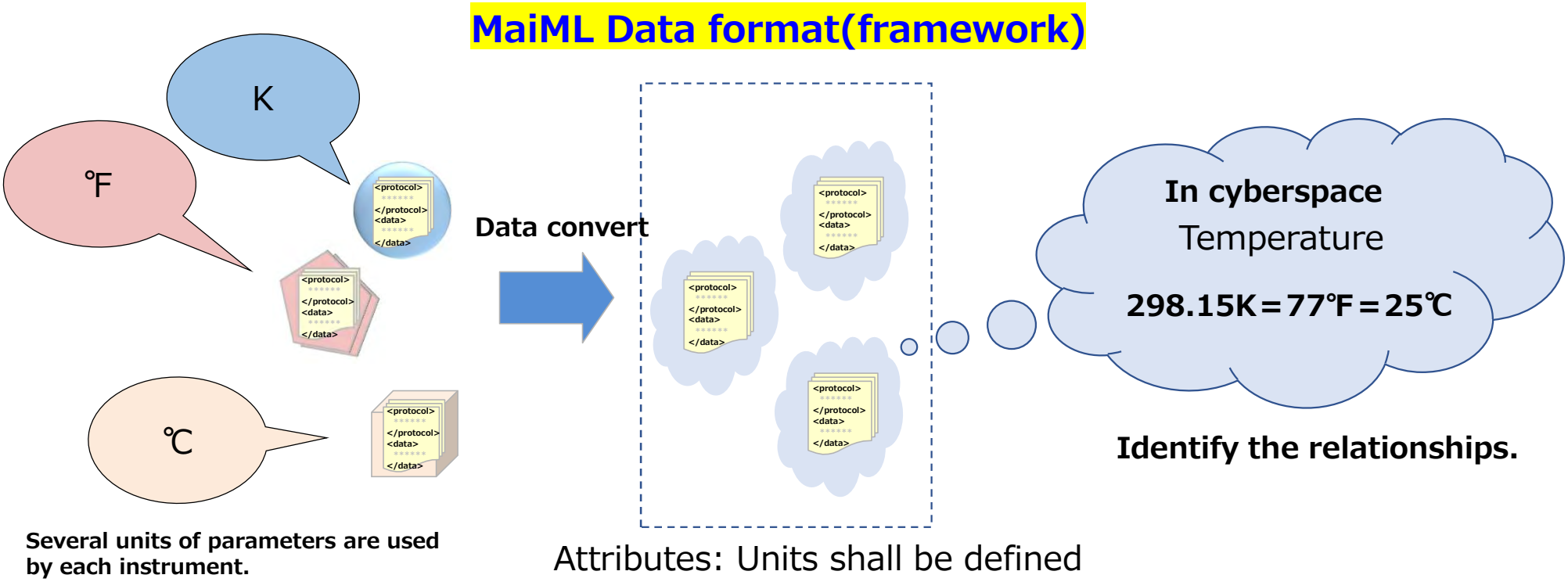
Key Advantages of MaiML: Generic Data Expression

- Data format and/or structure differ depending on manufacturers and/or instruments.
- Users can view or compare diverse data easily through an application.



Key Advantages of MaiML② :Generic Data Expression

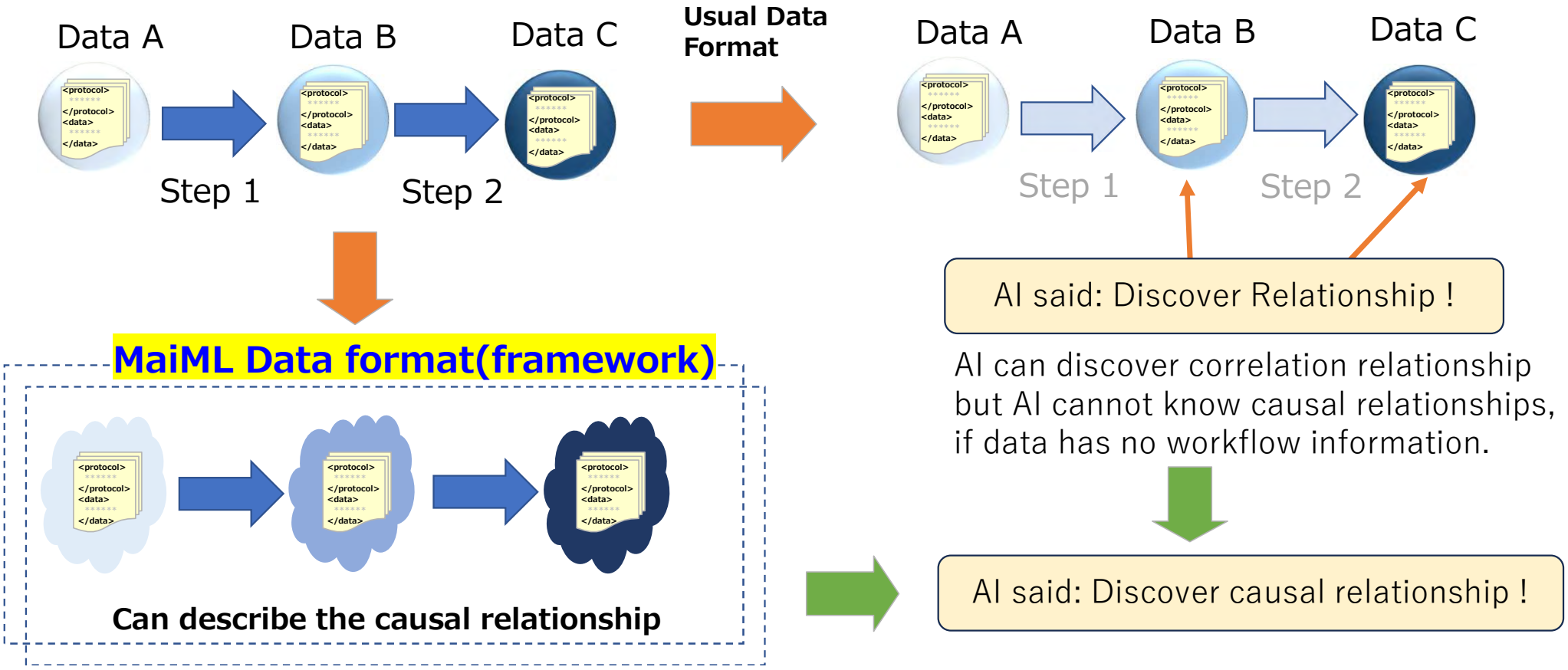
- Even different units of parameters by each instrument are used.
  → It can identify relationships between values, such as whether the same units are required.



Key Advantages of MaiML③: Workflow Significance

- If the causal relationships between data are described in the proposal, their meanings can be incorporated.

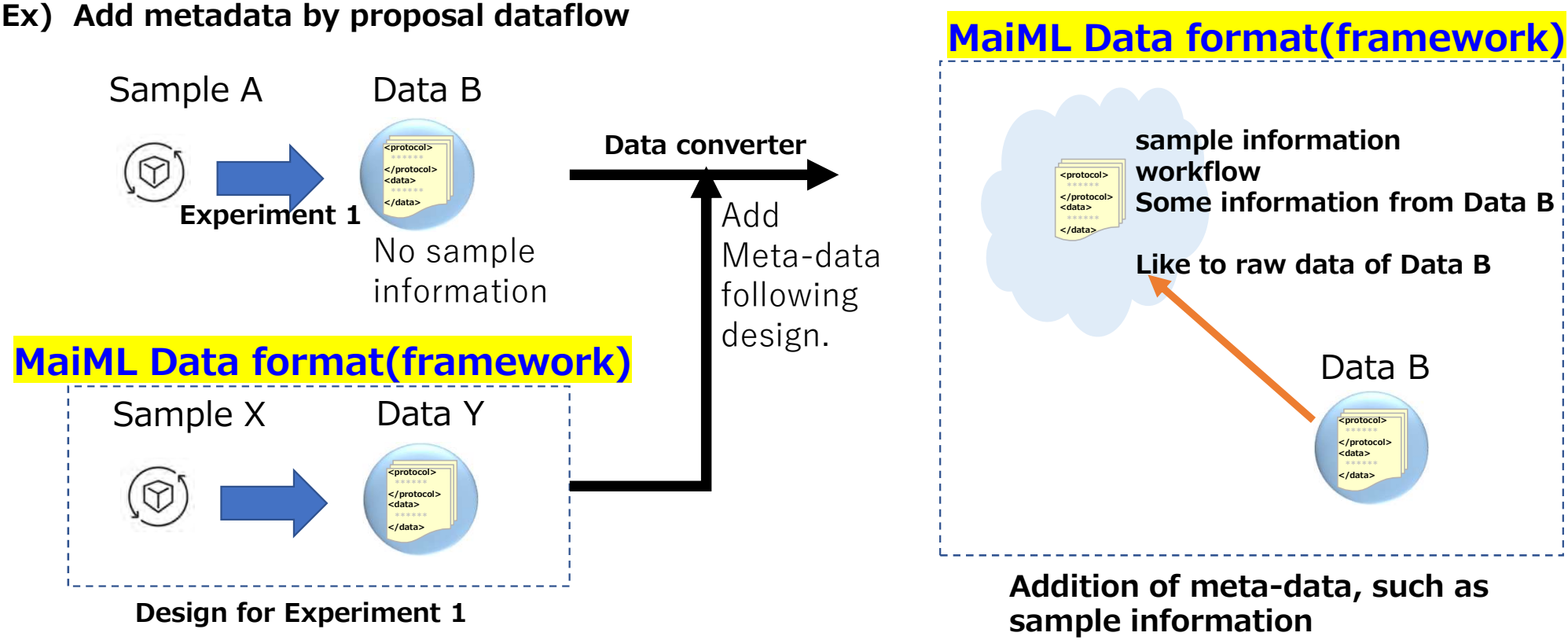
Ex) Two step data analysis



Key Advantages of MaiML④: Add metadata by flow

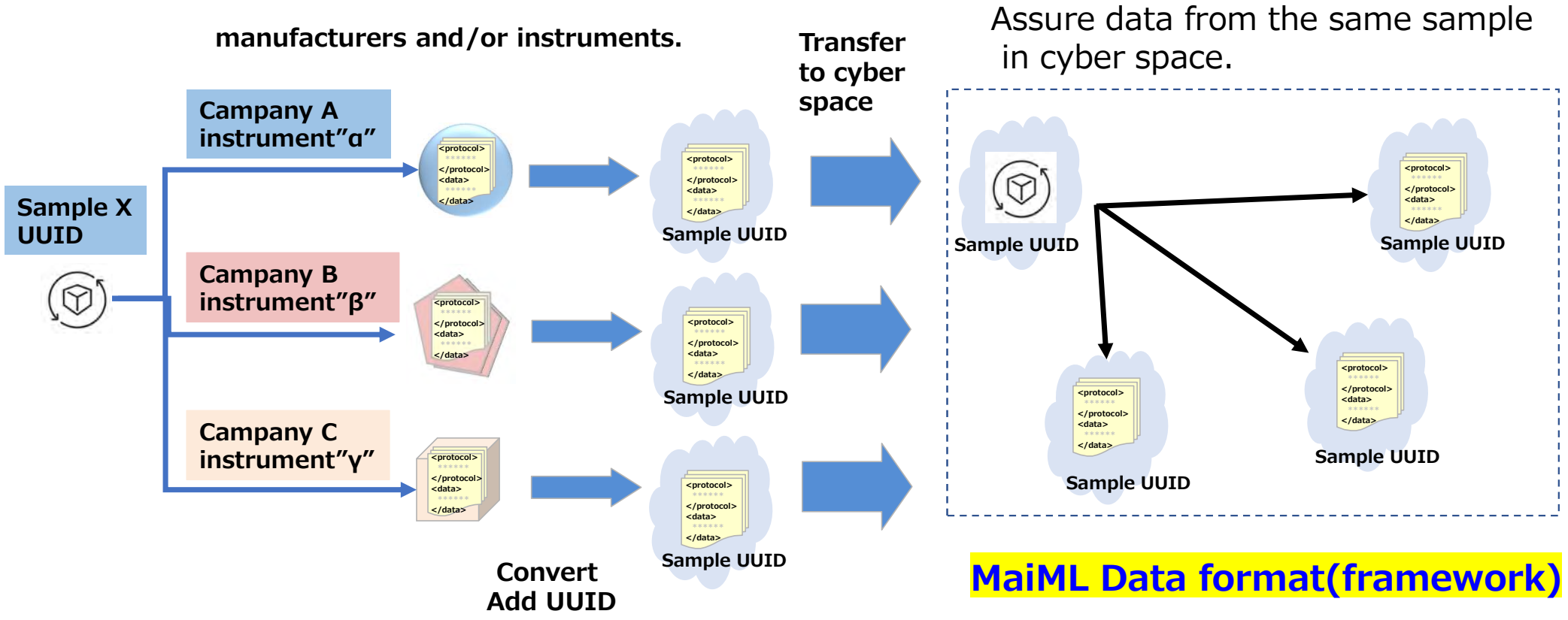
- Metadata can be added to workflows and templates designed for measurement and analysis.
- Data converter can add metadata following the workflow and Templates.

Ex) Add metadata by proposal dataflow



Key Advantages of MaiML⑤：Uniqueness Assurance

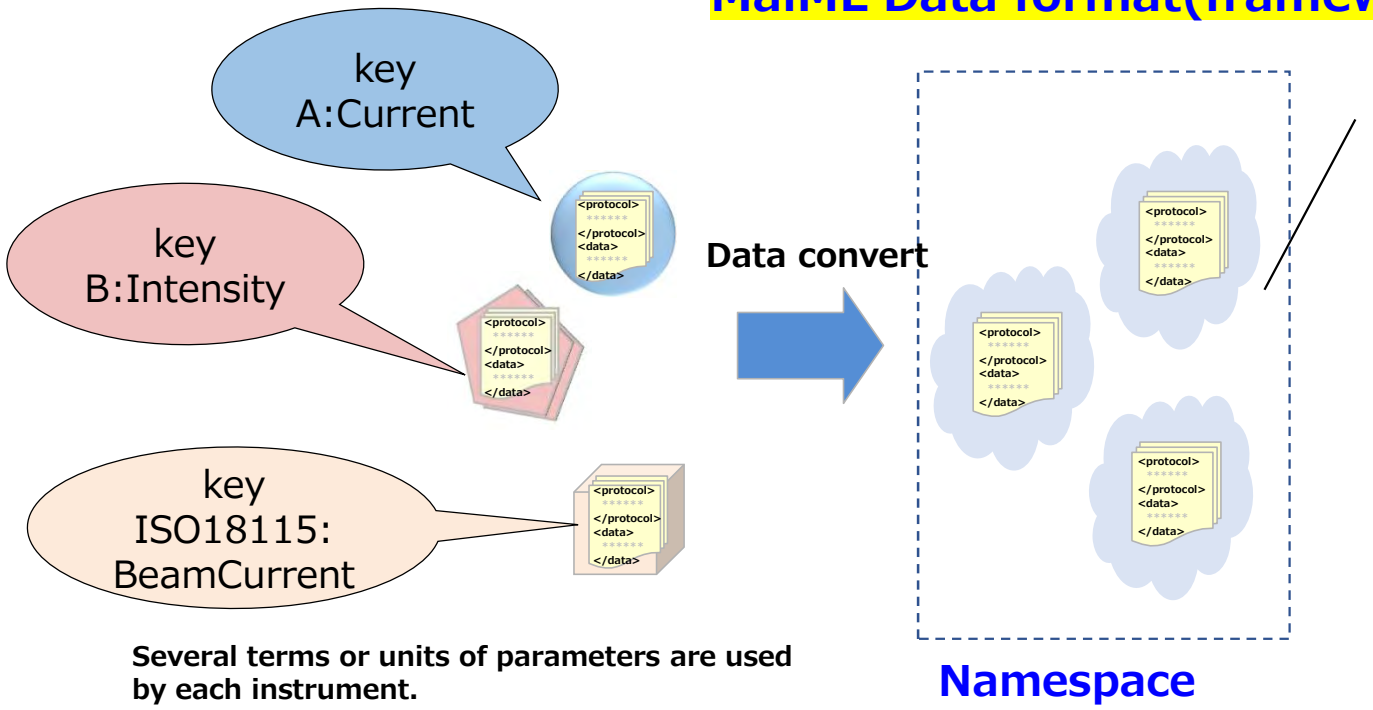
- In case the same sample is tested by different instruments, the uniqueness of the sample data must be assured. Ensuring this uniqueness allows AI to automatically identify the origin of the sample data.



Key Advantages of MaiML⑥：Namespace and ontology

- Several terms of parameters by each instrument are used.
  → Making the library to identify the relationships

MaiML Data format(framework)



Several terms or units of parameters are used by each instrument.

Direction library

Company A:
A:Current similar ISO18115:BeamCurent

Company B :
B:Intensity sameAs ISO18115:BeamCurrent

↓

A:Current similar B:Intensity

Define the relationships by Ontology

AI can identify that
A:Current is similar to B:Intensity but not the same.

Proposal for a comprehensive file format for the measurement and analysis of SCA data

- Proposal:
  - A comprehensive file format for measurement and analysis
    - guarantee of **independent availability** –

      - XML-compliant
      - Workflow/Process represented by Petri net
      - Class(template) for design of analysis and Instance for real data
      - Logging represented by XES
      - Detection of Falsification
      - Concealment

Proposed Items to be standardized

- **Comprehensive representation of measurement**
  - Document Uniqueness
  - Flow of measurement/analysis
    - Protocol
  - Real data/Instance
  - Traceability
- **Generic representation of data**
  - Type, meaning, value, uncertainty etc.
- **Modelling/design of measurement and analysis procedure**
  - Petri net representation
    - Material
    - Condition
    - Result
- **Detection of falsification and privatization**

Overviews: a comprehensive file format necessary for SCA
- the guarantee of independent availability/FAIR Alignment of MaiML -

The development of advanced materials requires the synthetic analysis of data obtained using various measurement equipment in the field of surface chemical analysis (TC201). Data format representing the whole of measurements and its pre- and post-processes is necessary for synthetic analysis. Thus, the format is required to guarantee independent availability as a platform related to FAIR+G principles.

FAIR+G Principle: Trust, Transparency, and Proper Access Control

F-Findable

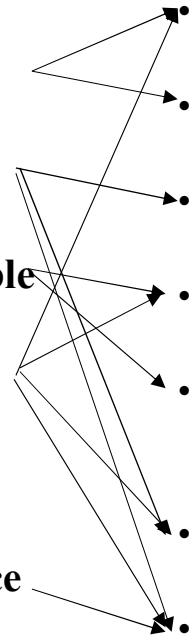
A-Accessible

I-Interoperable

R-Reusable

+

G-Governance



- **Independent Availability**: All of the necessary information in data documents alone.
  - Enable big data analysis such as AI and data science
- **Uniqueness assurance: Findable**
  - Universally Unique Identifier (UUID)s
- **Visualization**: Accessible
  - eXtensible Markup Language (XML)-compliant/Schema proposal
- **Reproducibility**: design/modelling of measurement and analysis procedure
  - Petri Net Markup Language (PNML)-compliant/Petri net-type representation
- **Generic data representation**
  - Attributes (data types), values, meaning of data
  - Inclusion of external data, various formats
- **Traceability**: Logging/provenance
  - Extensible Event Stream (XES)-compliant/Record of the time of measurement analysis
- **Open/Close strategies**
  - Concealment