UDM XML Schema Change Log

Changes in Version 6.0.0

XML Schema

1. Introduced controlled vocabularies (in separate *.xsd files) for
    - Countries (based on ISO 3166)
    - Units (based on 2019.09 Allotrope version of QUDT)
        - Amount of substance
        - Mass
        - Mass per volume
        - Molar concentration
        - Pressure or stress
        - Temperature
        - Time
        - Volume
    - Analytical methods (based on 2019.09 Allotrope Foundation Taxonomies)
    - Result types (based on 2019.09 Allotrope Foundation Taxonomies)
    - Reaction classes (based on the RXNO reaction ontology)
2. New entity: `ANALYTICAL_DATA` and related sub-elements: `EXPERIMENT`, `CREATION_DATE`, `SYSTEM`, `SUMMARY`, `DATA`, `DATA_URL`.
3. New entity: `ORGANISATION` containing `NAME` (required), `ADDRESS`, `COUNTRY`, `URL`, `EMAIL`, `PHONE`, `COMMENT`.
4. New entity: `ORGANISATIONS` to group `ORGANISATION`s.
5. Enhanced `AUTHOR` entity to include `NAME` (required), `EMAIL`, `PHONE`, `ORGANISATION`.
6. New entity: `AUTHORS` to group `AUTHOR`s.
7. `RESP_SCIENTIST` and `SCIENTIST` contain the same elements as `AUTHOR`.
8. New entity: `SAMPLE` containing `SAMPLE_ID` (required), `SAMPLE_REF`, `SAMPLE_MASS`, `BATCH_ID`, `AMOUNT`, `PURITY`, `BARCODE`, `ANALYTICAL_DATA`.

9. New entity: SAMPLES to group SAMPLEs.
10. The DATE field of COPYRIGHT renamed to YEAR, e.g. /UDM/LEGAL/COPYRIGHT/DATE becomes /UDM/LEGAL/COPYRIGHT/YEAR.
11. Type of /UDM/CITATIONS/CITATION/PATENT_PUB_DATE changed to xs:date.
12. Type of /UDM/CITATIONS/CITATION/YEAR changed to xs:positiveInteger.
13. MOLECULE can be specified inside REACTANT, PRODUCT, SOLVENT, CATALYST and REAGENT in addition or instead of the MOLECULES block.
14. Removed the ID attribute from REACTANT, PRODUCT, SOLVENT, CATALYST and REAGENT, use <MOLECULE MOD_ID="..." /> instead.
15. CITATION can be directly specified inside the MOLECULE and VARIATION entities either by referencing a citation from the CITATIONS element or providing CITATION subelements. It replaces the CIT_ID element in previous versions.
16. New placeholder entity EHS for future use (Environment and Health Safety data).
17. New placeholder entity ROUTES for future use (multistep reactions).

Data sets

Reaxys

1. Updated to version 6.0.0
2. Fixed RXNO_REACTION_TYPE values to be compatible with the RXNO ontology.

SPRESI

1. Conversion code updated to generate version 6.0.0
2. Data set updated to version 6.0.0

Changes in Version 5.0.1

XML Schema

Note

> XML Schema 1.0 has limited support for unambiguous language constructs and e.g. choices of elements cannot be fully restricted. One example is the types that should allow only one of the following subelements: min, max, min and max, exact. XML Schema allows only encoding one of the following two approximate scenarios: (a) one or more of the listed sub-elements or (b) zero or one of the listed subelements. We have assumed that it is safer to allow empty elements and incomplete information rather than multiple values of the same sub-element. This strategy is applied to elements like PH, PRESSURE, TEMPERATURE, TIME:

```
<!-- Allowed in the selected strategy, not allowed in the alternative one. --
>
<TEMPERATURE></TEMPERATURE>
<!-- Allowed in the selected strategy, not allowed in the alternative one. --
>
<TEMPERATURE>
  <max>100</max>
  <incr unit="deg_C/hour">5</incr>
</TEMPERATURE>
<!-- Not allowed in the selected strategy, allowed in the alternative one. --
>
<TEMPERATURE>
  <max>100</max>
  <min>90</min>
  <max>100</max>
  <incr unit="deg_C/hour">5</incr>
</TEMPERATURE>
```

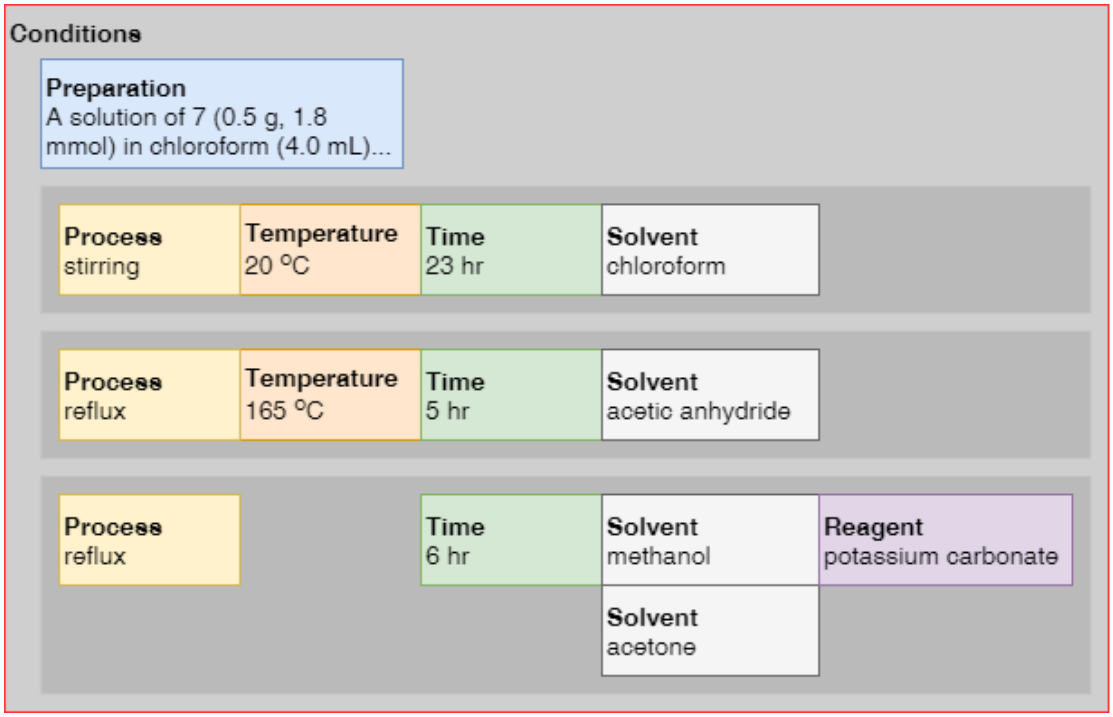
1. `CATALYSTS` renamed to `CATALYST`; removed hardcoded limits: eight `COMMENTS`, 20 `SAMPLE_ID`, 20 `SAMPLE_REF`, 20 `COMPOUND_NAME` per catalyst; Added new property: `ENANTIOMERIC_PURITY`.

2. Added new attributes for the `CLP` entity: `method`, `software`, `version`, `unit`

3. `COMMENTS` renamed to `COMMENT`.

4. `CONDITIONS`

   The reaction condition model has been significantly refactored to allow grouping of related conditions. In contrast to previous versions only single instance of the `CONDITIONS` element per variation is allowed. Furthermore, additional condition types have been introduced (`BUFFER_TYPE`, `BUFFER_CONCENTRATION`, `STIRRING`, `PROCESS`) as well as references to various types of agents (`REACTANT_ID`, `CATALYST_ID`, `SOLVENT_ID`, `REAGENT_ID`). Another change is that in this version the `PREPARATION` element can be used in two different locations: (1) at the very beginning of the `CONDITIONS` block and (2) inside `CONDITION_GROUP`. This allows representing more complex scenarios as the one illustrated below:



Conditions

Example:

```
<CONDITIONS>
  <PREPARATION>A solution of 7 (0.5 g, 1.8 mmol)...</PREPARATION>
  <CONDITION_GROUP>
    <TEMPERATURE>
      <min>20</min>
      <max>20</max>
    </TEMPERATURE>
    <TIME>
      <min>23</min>
      <max>23</max>
    </TIME>
  </CONDITION_GROUP>
  <CONDITION_GROUP>
    <TEMPERATURE>
      <min>165</min>
      <max>165</max>
    </TEMPERATURE>
    <TIME>
      <min>5</min>
      <max>5</max>
    </TIME>
  </CONDITION_GROUP>
  <CONDITION_GROUP>
    <TIME>
      <min>6</min>
```

```
      <max>6</max>
    </TIME>
  </CONDITION_GROUP>
</CONDITIONS>
```

5. IDENTIFIERS renamed to IDENTIFIER.

6. New element LEGAL

   New, optional element to represent license and copyright information, to be used before CITATIONS.

   Example:

```
<LEGAL>
  <PRODUCER>Jarek Tomczak</PRODUCER>
  <TITLE>Sample UDM 5.0.0 dataset</TITLE>
  <LICENSE href="https://creativecommons.org/licenses/by-nd/4.0">
    Creative Commons Attribution-NoDerivatives 4.0 International
    (CC BY-ND 4.0)
  </LICENSE>
  <COPYRIGHT href="http://pistoiaalliance.org/projects/udm">
    <TEXT>Copyright (c) 2018 Pistoia Alliance</TEXT>
    <OWNER>Pistoia Alliance</OWNER>
    <DATE>2018</DATE>
  </COPYRIGHT>
</LEGAL>
```

7. LINKS renamed to LINK.

8. Removed the limit of maximum eight COMMENTS per metabolite in the METABOLITES element.

9. Extensions to MOLSTRUCTURE

   - New format attribute with the following allowed values: cdxml, inchi, molfile, smiles and wiswesser. If not specified the default molfile value is assumed.

   - Multiple instances of the MOLSTRUCTURE are now allowed so that different representations of a molecular structure can be used at the same time.

   Example:

```
<MOLECULE ID="1">
  <MOLSTRUCTURE format="molfile"><![CDATA[
  ACCLDraw10181722532D

 10 10  0  0  0  0  0  0  0  0999 V2000
    3.5382   -1.1166    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.4918   -1.1166    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.5150   -1.7074    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
```

```
    2.5150   -2.8888    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.5574   -2.8878    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.5342   -3.4785    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.5342   -4.6600    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5175   -5.2495    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4919   -4.6653    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4919   -3.4791    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  3  1  1  0  0  0  0
  3  2  2  0  0  0  0
  4  3  1  0  0  0  0
  4 10  1  0  0  0  0
  6  4  2  0  0  0  0
  6  5  1  0  0  0  0
  7  6  1  0  0  0  0
  8  7  2  0  0  0  0
  9  8  1  0  0  0  0
 10  9  2  0  0  0  0
M  END
]]></MOLSTRUCTURE>
  <MOLSTRUCTURE format="smiles">OC(=O)c1ccccc1O</MOLSTRUCTURE>
  <!-- Other elements... -->
</MOLECULE>
```

10. Extended support for `PH` values. The following combinations of sub-elements of `PH` are now supported:

    - `min` only
    - `max` only
    - `min` and `max`
    - `exact`

    Example:

```
<PH>
  <max>10.0</max>
</PH>

<PH>
  <min>7.0</min>
  <max>10.0</max>
</PH>

<PH>
  <exact>7.0</exact>
</PH>
```

11. Enhanced `PREPARATION` entity:

- Type changed to allow any content: text string, XML elements, mixed and CDATA.
- Added optional `format` attribute, default value is `text`.
- Allowed multiple instances of the preparation section to allow multiple representation of the same content.

Example:

```
<PREPARATION format="html">
  Sat. aqueous Na<sub>2</sub>S<sub>2</sub>O<sub>4</sub> (20 mL) was added
to a solution of 10 (7.6 mg, 0.020 mmol) in ...
</PREPARATION>
<PREPARATION format="text">
  Sat. aqueous Na2S2O4 (20 mL) was added to a solution of 10 (7.6 mg,
0.020 mmol) in ...
</PREPARATION>
<PREPARATION format="pdf">
  <![CDATA[%PDF-1.7
%
1 0 obj
<</Type/Catalog/Pages 2 0 R/Lang(en-GB) /StructTreeRoot 10 0
R/MarkInfo<</Marked true>>/Metadata 20 0 R/ViewerPreferences 21 0 R>>
endobj
...
]]>
</PREPARATION>
<PREPARATION format="robot">
  <STEP action="add">
    <COMPOUND>Sodium dithionite</COMPOUND>
    <VOLUME unit="mL">20</VOLUME>
    ...
  </STEP>
</PREPARATION>
```

12. Changed `PRESSURE` entity:
    - The `unit` attribute is now specified for the `PRESSURE` element rather than `min` and `max` as in 4.0.0.
    - The following combinations of sub-elements of PRESSURE are now supported:
      - `min` only
      - `max` only
      - `min` and `max`
      - `exact`

13. PRODUCTS renamed to PRODUCT; removed hardcoded limits: eight COMMENTS, 20 SAMPLE_ID, 20 SAMPLE_REF, 20 COMPOUND_NAME per product; added new properties: EXPECTED_AMOUNT, EXPECTED_MASS, SAMPLE_MASS.

14. New PROPERTY element to store molecular properties within a MOLECULE. Attributes: name, method, software, version and unit.

15. New PSA attributes in addition to unit introduced in 4.0.0: method, software, version.

16. REACTANTS renamed to REACTANT; removed hardcoded limits: eight COMMENTS, 20 SAMPLE_ID, 20 SAMPLE_REF, 20 COMPOUND_NAME per reactant. Added new properties: DENSITY, ENANTIOMERIC_PURITY, LOADING, MOLARITY, PERCENT_WEIGHT, SAMPLE_MASS.

17. REAGENTS renamed to REAGENT; removed hardcoded limits: eight COMMENTS, 20 SAMPLE_ID, 20 SAMPLE_REF, 20 COMPOUND_NAME per reagent; added new properties: MOLARITY and PURITY.

18. Updated REACTION_MOLARITY:

    - The unit attribute is now specified for the REACTION_MOLARITY element rather than min and max as in 4.0.0.

    - The following combinations of sub-elements of REACTION_MOLARITY are now supported:

        - min only
        - max only
        - min and max
        - exact

19. Enhanced RXNSTRUCTURE - new format attribute with the following allowed values: cdxml, rinchi, rsmiles and rxn. If not specified the default rxn value is assumed.

20. New SECTION element

    New element allowing storing publisher-specific data in predefined places of the UDM hierarchy: CITATION, MOLECULE, REACTION, VARIATIONS, REACTANT, PRODUCT, SOLVENT, CATALYST, REAGENTS, CONDITIONS. The contents of the SECTION element can contain any data and it will be ignored when validated against the UDM schema, but a derived schema can be created which will extend the UDM schema and provide additional restriction on the content of the SECTION elements.

    Example: the reaction PRODUCT element may contain analytical data: xml <PRODUCT> ... <SECTION type="analytical_data"> <!-- Custom, publisher-specific data. --> <Spectrum type="IR" id="423"> <![CDATA[...]]> </Spectrum> </SECTION> </PRODUCT>

The SECTION element will be processed without validation when using the standard UDM schema, but the example XML schema below provides additional validation by enforcing the SECTION element to contain Spectrum sub-elements which may have type and id attributes:

```
<?xml version="1.0" encoding="utf-8" ?>
<!-- Sample udm_spectra.xsd schema. -->
<xs:schema elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:redefine schemaLocation="udm_pa_5_0_0_draft.xsd">
    <xs:complexType name="sectionData">
      <xs:complexContent>
        <xs:restriction base="sectionData">
          <xs:sequence>
            <xs:element name="Spectrum">
              <xs:complexType>
                <xs:simpleContent>
                  <xs:extension base="xs:string">
                    <xs:attribute name="type">
                      <xs:simpleType>
                        <xs:restriction base="xs:string">
                          <xs:enumeration value="IR" />
                          <xs:enumeration value="NMR" />
                          <xs:enumeration value="MS" />
                          <xs:enumeration value="UV" />
                        </xs:restriction>
                      </xs:simpleType>
                    </xs:attribute>
                    <xs:attribute name="id" type="xs:nonNegativeInteger"
/>
                  </xs:extension>
                </xs:simpleContent>
              </xs:complexType>
            </xs:element>
          </xs:sequence>
          <xs:attribute name="type" type="xs:string" />
        </xs:restriction>
      </xs:complexContent>
    </xs:complexType>
  </xs:redefine>
</xs:schema>
```

21. SOLVENTS renamed to SOLVENT.
    - Removed hardcoded limits: eight COMMENTS, 20 SAMPLE_ID, 20 SAMPLE_REF, 20 COMPOUND_NAME per solvent.
    - Added new properties: DENSITY, EQUIVALENTS.
22. Removed element STAGES - use CONDITIONS and CONDITION_GROUPS instead.

23. SYNONYMS renamed to SYNONYM.

24. Updated TEMPERATURE:

    – The unit attribute is now specified for the TEMPERATURE element rather than min and max as in 4.0.0.

    – The following combinations of sub-elements of TEMPERATURE are now supported:

        • min only

        • max only

        • min and max

        • min, max and incr

        • exact

    Examples:

```
<TEMPERATURE>
  <min>90</min>
</TEMPERATURE>

<TEMPERATURE unit="degree_C">
  <min>90</min>
  <max>100</max>
  <incr unit="degree_C/hour">5</incr>
</TEMPERATURE>

<TEMPERATURE unit="degree_C">
  <exact>93.2</exact>
</TEMPERATURE>
TIME
```

25. Updated TIME:

    – The unit attribute is now specified for the TIME element rather than min and max as in 4.0.0.

    – The following combinations of sub-elements of TIME are now supported:

        • min only

        • max only

        • min and max

        • exact

26. Updated TOTAL_VOLUME:
    - The unit attribute is now specified for the TOTAL_VOLUME element rather than min and max as in 4.0.0.
    - The following combinations of sub-elements of TOTAL_VOLUME are now supported:
        - min only
        - max only
        - min and max
        - exact
27. Removed UDM_VERSION_PA element

    This element was introduced in version 4.0 to represent the actual version of the Pistoia Alliance UDM and do not break compatibility with the original Elsevier version (with hardcoded UDM_VERSION).

28. Updated UDM_VERSION

    Removed the BUILD component and thus simplifying versioning to MAJOR.MINOR.REVISION. Changed allowed version values to 5.x.x where x are integers starting from 0.

29. VARIATIONS renamed to VARIATION; added new outcome attribute that can be used to store arbitrary assessment whether the reaction was successful or failed.

Other

1. UDM code hosted on GitHub: https://github.com/PistoiaAlliance/UDM

Changes introduced in 4.0.0

First public release, fully compatible with the original version 3.6.0.112 donated by Elsevier.

1. Added explicit unit of measure to a number of entities.
2. Added documentation for key entities.
3. Included sample data sets (SPRESI, Reaxys)
4. Included conversion tool from SPRESI RD file to UDM
5. Numbered as 4.0.0 to avoid future conflicts with existing Elsevier versions