XML Extraction

Description

The handler extracts the data table from the column of the input table in the cells of which the XML documents are kept in the string form. All documents from the column must comply with the connected XSD schema.

Ports

Input

- — Connect to XSD schema source allows for connection of the configured earlier connection to the XSD schemas files.
- — Control variables (variables), optional.
- — Data source: the data table that contains the cells with the XML documents.

Output

- — Data set: the table of data extracted from the XML documents.
  - There are the following fields the availability of which is set by a user:
    - The field identifier enables to identify from which document a particular string has been extracted.
- — Errors and exit codes — a table of extraction logs that has the following structure:
  - Required fields:
    - The exit code contains the values of extraction codes for each document.
  - There are the following fields the availability of which is set by a user:
    - The field identifier enables to identify the document with which a record is connected in the log table.
    - Error description: the description of exit codes.

Wizard

Step 1. Parameters of XML Extraction

- **XML column**: selection of the column with the XML documents.
- **XML Document Identification**: availability of the identifier column in the extracted table and log table is defined in this area:
  - Use identifier: the checkbox enables identifiers. Then it is required to define the type of identifiers using the radio button:
    - Number of the input data set row: the number of the string which contains a document. 0 identifier matches the first string.
    - The column value: values of the selected column of the input table will be used as identifiers.
- **Halt on errors**: this checkbox enables to cancel extraction when the first error is defined.
- **Unescape XML entities** is applied if the documents are extracted from another XML document in which they have been escaped. When this checkbox is not available, extraction of the document that starts, for example, from such notice as `<`

```
&lt;?xml version=&quot;1.0&quot; encoding=&quot;UTF-8&quot;?&gt
```

will cause an error, and when this checkbox is selected, the handler will interpret such notice as the correct one:

```
<?xml version=«1.0» encoding=«UTF-8»?>
```

- **Provide description of error codes**: "Error description" column will be added to the table with logs.

Step 2. Imported Fields

- **Connection**: the state of connection to the XSD schema.
- **Root element**: selection of the XSD schema root element according to which data will be extracted. The area displaying the hierarchic structure of the selected root element is located below. It is required to select the checkboxes for the elements that form the fields of the extracted table.
- **Default time zone**: the *Date/Time* type values are transformed when extracting from XML according to the following algorithm:
  - If the time zone is specified for the value, it is transformed to the current time zone of the Megaladata server.
  - If the time zone is not specified for the value, the following conditions are met:

    - If *Default time zone* is specified, the value is transformed to the current time zone of the Megaladata server based on the assumption that it has the time zone specified in the *Default time zone* parameter.
    - If *Default time zone* is not specified, the value is not changed.
- **Repeat single values**: if there are a single item and nested sequence included into a sequence, the single item values will be repeated for each row formed from the nested sequence.
- **Check for strict XSD validation**: when this checkbox is selected, XML is validated for strict compliance with the XSD schema. Otherwise, the check is performed in the simplified accelerated mode. The simplified check procedure is performed faster, however, it does not secure correct data extraction from XML if there are undetected validation errors.
- **Generate composite field captions**: if this checkbox has been selected, hierarchy relative to the root element will be displayed in each field caption. For example, complex "ROOT|PERSON|ACCES-LEVEL" caption will be generated for the extracted table field that contains the data of "ACCESS-LEVEL" attribute of "PERSON" element from the root element of "ROOT" document. When there is no such checkbox, the field will get the following name: "ACCESS-LEVEL".