Automatic Interpretation of Complex UDC Numbers: Towards Support for Library Systems

Attila Piros

University of Debrecen, Hungary

UDC Seminar, Lisbon, 29-30 October 2015

Goal

- Supporting software systems to utilize **Universal Decimal Classification** to retrieve information effectively:
  - **Find a way to represent each language units of** simple and complex **UDC numbers** in an easy processable format by keeping all of the information stored in the numbers
  - **Create an algorithm and implement a reference application to interpret UDC numbers** by automatic means
  - **Develop and implement conversion methods** to different formats

The representation of UDC numbers

- The usual structure of a **‘simple’** UDC number:

  [main table number/range][special auxiliaries][dependent common auxiliaries][independent common auxiliaries (containing numbers/ranges, special auxiliaries and mayhap operations)]

- **Compound** (or ‘complex‘) UDC numbers are built from ‘simple’ numbers by using auxiliary signs
- **Subgrouping** can be used to clarify the order or modify compounds by auxiliaries

The representation of UDC numbers

- After investigating the definitions of the **operations** the following **precedence order** can be defined:
  - + Coordination
  - : Simple relation
  - :: Order-fixing
    - [ ] Subgrouping
  - ' Synthesis (within special auxiliaries)
- Every UDC number can be represented with a **tree**
- It is possible to define a **schema definition** to determine the exact format to describe the numbers in **XML**

The representation of UDC numbers

- The **complex types** of the XSD describe the possible elements of UDC numbers, e.g. **schedule numbers**:

```
<xsd:complexType name="special_auxiliary_number_hyphen">
  <xsd:complexContent>
    <xsd:restriction base="udc:special_auxiliary_number">
    <xsd:attribute name="number1“
  type="udc:special_auxiliary_number_hyphen_string“
  use="required"/>
    <xsd:attribute name="number2"
  type="udc:special_auxiliary_number_hyphen_string"
  use="optional"/>
    </xsd:restriction>
  </xsd:complexContent>
</xsd:complexType>
```

The representation of UDC numbers

- A common **auxiliary sign** (operation) can be described by a **complex type** containing its possible operands:

```
<xsd:complexType name="main_table_relation">
<xsd:sequence>
   <xsd:choice minOccurs="2" maxOccurs="unbounded">
    <xsd:element name="main_table_number"
type="udc:main_table_number" minOccurs="1" maxOccurs="1"/>
    <xsd:element name="main_table_synthesis"
type="udc:main_table_synthesis" minOccurs="1" maxOccurs="1"/>
    <xsd:element name="main_table_subgrouping"
type="udc:main_table_subgrouping" minOccurs="1" maxOccurs="1"/>
    <xsd:element name="main_table_orderfixing"
type="udc:main_table_orderfixing" minOccurs="1" maxOccurs="1"/>
   </xsd:choice>
 </xsd:sequence>
</xsd:complexType>
```

The representation of UDC numbers

- **Simple types** have been introduced for validation purposes:

```
<xsd:simpleType
   name="special_auxiliary_number_hyphen_string">
   <xsd:restriction base="xsd:string">
      <xsd:pattern value="-\d(\d{0,1}|(\d{2}(\.[1-9]\d{2})*(\.[1-
   9]\d{0,2})?))"/>
      <xsd:minLength value="2"/>
   </xsd:restriction>
</xsd:simpleType>
```

The automatic interpretation of UDC numbers

- Converting UDC numbers **manually** to a complex format such as that mentioned earlier is an **unrealistic** expectation
- The **existing records** should also be processed and converted
- The **UDC number** itself is a **common and stable element** of the varying formats

The automatic interpretation of UDC numbers

- Has been researched **for about 50 years**
- A comprehensive research was conducted by **Gerhard Riesthuis** (Zoeken met woorden, 1998)
- In the course of this research, a **new algorithm** has been created, which is better suited to the XML schema and the principles which will be explained on the next slide

The automatic interpretation of UDC numbers

- The algorithm must recognize those numbers which **keep to the rules** for synthesizing UDC numbers
- The algorithm must retain **all of the information** stored by the number, containing all of its parts and the information pertaining to their context and role
- The parsing method must be **fully syntactic** as far as is possible
- The process must be **fully automated**

The automatic interpretation of UDC numbers

- **Online availability** and providing outputs in **different formats** are also important expectations
- The software is **available online** for testing purposes on the following URL: http://interpreter-eto.rhcloud.com/

The automatic interpretation of UDC numbers

- 821.111SHAK7ROM.03=112.2

```
<ns:udc_concept xmlns:ns="http://library.inf.unideb.edu/udc/xml"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" udc_edition="2005"
   notation="821.111SHAK7ROM.03=112.2">
   <ns:description xml:lang="EN">
        Shakespeare: Romeo and Juliet (translated into German)
   </ns:description>
   <ns:main_table_number number1="821.111">
        <ns:special_auxiliary xsi:type="ns:special_auxiliary_number_numerical" umber1="7"/>
        <ns:special_auxiliary xsi:type="ns:special_auxiliary_number_pointnought" number1=".03"/>
        <ns:alphabetical_specification order="1" text="SHAK" standard=""/>
        <ns:alphabetical_specification order="2" text="ROM" standard=""/>
        <ns:common_auxiliary_independent xsi:type="ns:common_auxiliary_of_language" order="1">
             <ns:common_auxiliary_of_language_number number1="=112.2"/>
        </ns:common_auxiliary_independent>
   </ns:main_table_number>
</ns:udc_concept>
```

The automatic interpretation of UDC numbers

- 061.1(100)::[54+66]

```
<ns:udc_concept xmlns:ns="http://library.inf.unideb.edu/udc/xml" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" udc_edition="2005" notation="061.1(100)::[54+66]">
    <ns:description xml:lang="EN">
        IUPAC - International Union of Pure and Applied Chemistry
    </ns:description>
    <ns:main_table_orderfixing>
        <ns:main_table_number number1="061.1" order="1">
            <ns:common_auxiliary_independent xsi:type="ns:common_auxiliary_of_place">
                <ns:common_auxiliary_of_place_number number1="(100)"/>
            </ns:common_auxiliary_independent>
        </ns:main_table_number>
        <ns:main_table_subgrouping order="2">
            <ns:main_table_addition>
                <ns:main_table_number number1="54"/>
                <ns:main_table_number number1="66"/>
            </ns:main_table_addition>
        </ns:main_table_subgrouping>
    </ns:main_table_orderfixing>
</ns:udc_concept>
```

The conversion of the results

- **KWOC-outputs** in JSON and HTML (have been available since August):
  - 394.4:[92(100+437):329(437).15(091)+327.32(100)], JSON
  - 394.4:[92(100+437):329(437).15(091)+327.32(100)], HTML
  - 510.2/.6, HTML

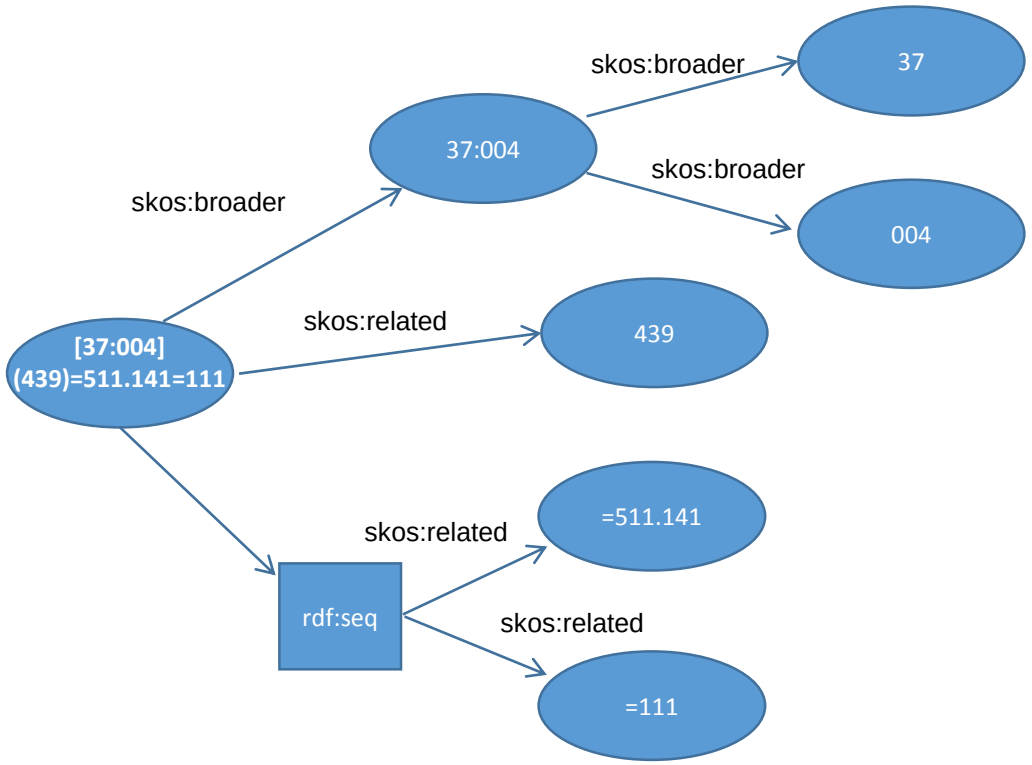
The conversion of the results

- **Standardized numbers** (has been available since October):
  - 622(437.3)333/.336-022.316=162.3(043)
  - 659.131.7.03:070.485
  - 821.111(73)-32=511.141(082)
  - 821.111(73)-32=511.141(082) (by keeping citation order)

The conversion of the results

- Supporting the further formats are **planned** and under design (but have not been published yet)
  - **RDF/SKOS**
  - **MARC**

The conversion of the results



The conversion of the results

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:udcpcs="http://interpreter-
eto.rhcloud.com/rdf/rdf-schema#">
 <skos:ConceptScheme rdf:about="http://interpreter-eto.rhcloud.com/rdf/rdf-schema">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#ConceptScheme"/>
  <dcterms:title>Complec UDC numbers</dcterms:title>
  <skos:note/>
 </skos:ConceptScheme>
 <skos:Concept rdf:about="http://interpreter-
eto.rhcloud.com/rdf/db/F0T0T4D3T7CJ4T3T9G5T1T1T1T4T1G1T1T1C">
  <skos:notation
rdf:datatype="http://udcdata.info/UDCnotation">[37:004](439)=511.141=111</skos:notation>
  <skos:prefLabel xml:lang="en">Computers in the education in Hungary (in English and
Hungarian)</skos:prefLabel>
  <skos:broader rdf:resource="http://interpreter-eto.rhcloud.com/rdf/db/0T0T4D3T7C"
udcpcs:commonAuxiliary="http://udcdata.info/000007"/>
  <skos:related rdf:resource="http://udcdata.info/003965"/>
  <rdf:seq udcpcs:commonAuxiliary="http://udcdata.info/000008">
   <rdf:li>
    <skos:related rdf:resource="http://udcdata.info/000789"/>
   </rdf:li>
   <rdf:li>
    <skos:related rdf:resource="http://udcdata.info/000028"/>
   </rdf:li>
  </rdf:seq>
 </skos:Concept>
</rdf:RDF>
```

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:udcpcs="http://interpreter-eto.rhcloud.com/rdf/rdf-schema#">
 <skos:ConceptScheme rdf:about="http://interpreter-
eto.rhcloud.com/rdf/rdf-schema">
  <rdf:type
rdf:resource="http://www.w3.org/2004/02/skos/core#ConceptScheme
"/>
  <dcterms:title>Complex UDC numbers</dcterms:title>
  <skos:note/>
 </skos:ConceptScheme>
 <skos:Concept rdf:about="http://interpreter-
eto.rhcloud.com/rdf/db/0T0T4D3T7C">
  <skos:notation
rdf:datatype="http://udcdata.info/UDCnotation">37:004</skos:notatio
n>
  <skos:prefLabel xml:lang="en">Computers in the
education</skos:prefLabel>
  <skos:broader rdf:resource="http://udcdata.info/024974"
udcpcs:commonAuxiliary="http://udcdata.info/000005"/>
  <skos:broader rdf:resource="http://udcdata.info/013566"
udcpcs:commonAuxiliary="http://udcdata.info/000005"/>
 </skos:Concept>
</rdf:RDF>
```

The conversion of the results

- **MARC21**

  084 8#$audc$bUniversal Decimal Classification$dBIP 0017-1 : 2005$eeng

  153 ##$a796.332$c796.333.4$e796.33$f796.333$hSports. Games. Physical exercises$hBall games$hBall games in which the ball is played with foot and hand$jAssociation and Rugby footballs

  353 ##$a796.333.3$iRugby union football

- **XML**

```
<ns:udc_concept xmlns:ns="http://library.inf.unideb.edu/udc/xml"
        udc_edition="2005"  notation="796.332/796.333.4">
 <ns:description xml:lang="EN"/>
 <ns:main_table_number number1="796.332" number2="796.333.4"/>
</ns:udc_concept>
```

The conversion of the results

- A software may be able to **automatically**
  - Recognize **hierarchical** (BT/NT) **relationships** between a **number and its parts**
  - Recognize **associative** (RT) **relationships** between a number and **its parts** (e.g. common auxiliaries)
  - Provide **suggestions** on **associative relationships** between **similar numbers**
- Utilizing **MRF** may be required to
  - Identify the exact **broader concepts** of schedule numbers
  - Identify the exact **special auxiliaries**
  - Recognize hierarchical relationships by **breaking down ranges**
  - Provide **suggestions** on associative relationships based upon the **references in the schedules**
- **Human interaction** may be required to
  - Find additional **associative relationships** between complex concepts

Thank you for your attention