

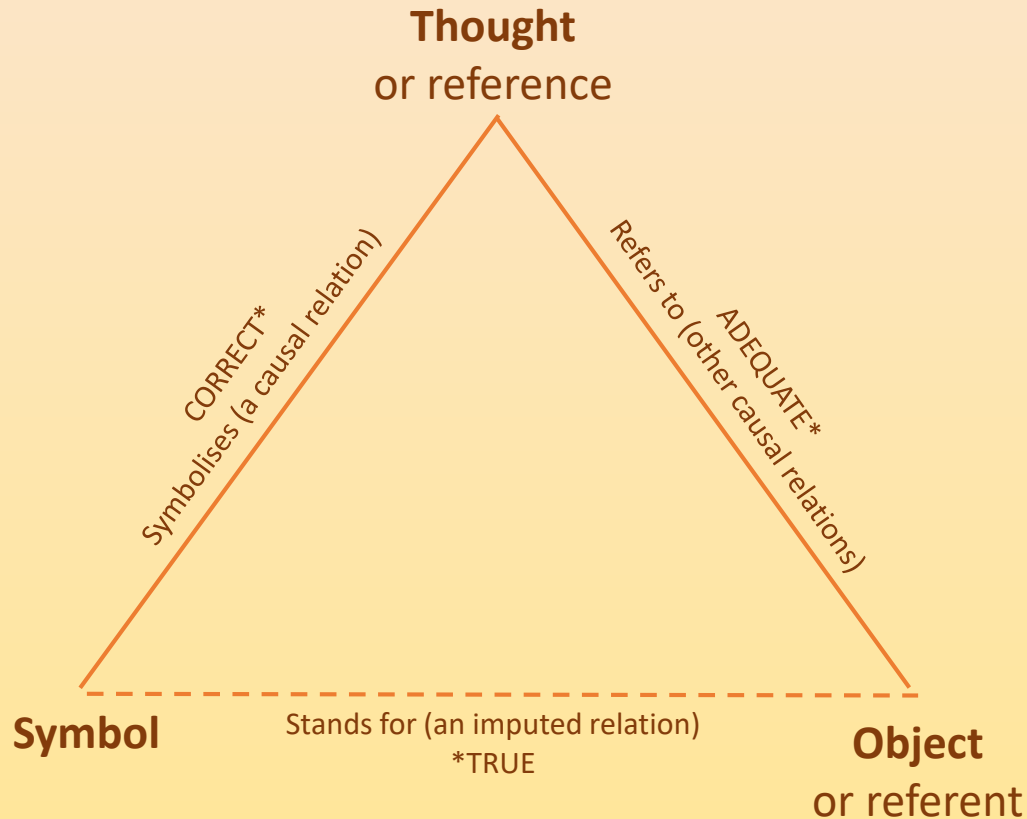
# The thought behind the symbol

**About the automatic interpretation and  
representation of UDC numbers**

**Attila Piros**

University of Debrecen, Hungary

# The thought behind the symbol



Semiotic triangle from "The Meaning of Meaning"  
by Charles Kay Ogden and Ivor Armstrong Richards (1946).

## In **classificatory subject metadata**:

- The **object** means the object of the description
- The **thought** is the resume of the object
- The **symbol** is the designation of the thought by means of a classification system

The decisions about the relevance require extracting the thought from the symbol.

# Fundamental structures in classifications

- **Enumerative schemes** (e.g. LCC, DDC)
- **Faceted schemes** (e.g. CC, BC2)
- **Analytico-synthetic schemes** (e.g. UDC, BC1). (Broughton 2015)
  - Has **systematic schedules** with rich concept hierarchy
  - Has (common and special) **auxiliary schedules**
  - Has facilities for expressing **complex and agglomerated subjects**

# UDC as an analytico-synthetic classification

- **Compound subjects:** concepts that are created by specifying basic subjects with facets. (Ranganathan 1967)
  - Special auxiliaries. Common auxiliaries. Table 1c/1h.
- **Complex subjects:** concepts that are created by joining two or more subjects on the bases of some relation. (Ranganathan 1967)
  - Relation (:.) and order-fixing (::). Table 1b.
- **Agglomerated basic subjects:** concepts that are built by the "collecting together of entities into larger masses without cohesion among the components". (Neelameghan 1973)
  - Coordination and consecutive extension. Table 1a.

# Compositionality

- The **idea** that "The meaning of a complex expression is determined by its structure and the meanings of its constituents."
- **Examples** of the semantic effects
  - Special and common auxiliary schedules
    - **512.54.03:531.111.5-3** – The relationship between elementary theory of various classes of groups and three-dimensional symmetry of space and time.
    - **512.54-3:531.111.5.03** – Computation technics in group theory in (investigating) the causes of phenomena in symmetry of space and time.
    - **341.232.3(519)::330.34(510)** – Economic aid to China by Korea
    - **341.232.3(510)::330.34(519)** – Economic aid to Korea by China
  - Auxiliary signs
    - **21+29** – Prehistoric religions and modern spiritual movements
    - **21/29** – Religious systems. Religions and faiths
    - **21:29** – The relationship between prehistoric religions and modern spiritual movements
    - **339.9(44:450)** – International economical relationships between France and Italy
    - **339(44+450)** – The economies of France and Italy
  - Citation order (Robinson 2003)
    - **94(410)"20"(051)** – Periodicals about the history of Britain in the 21<sup>st</sup> Century
    - **94(410)(051)"20"** – Periodicals from the 21<sup>st</sup> Century about the history of Britain
    - **94"20"(051)(410)** – British periodicals about the history of the 21<sup>st</sup> Century

# The goal of my current research

To support applications to **employ UDC**

To support the **decisions about the relevance** of UDC based subject metadata

To support the **analysis of UDC classmarks**

- To design a **format to store UDC numbers** by retaining both their elements and their inner structure
- To design and implement an algorithm that is able to **convert UDC numbers** into such a format by automatic means

# Requirements for the format

The **representation** of UDC numbers

- Has to describe the **whole syntactic structure** of the numbers
- Has to respect all **UDC rules** by taking the **different editions** into account
- Has to be a standard, **platform-independent** format

The **XML language** was chosen as the basis of the format, because it is:

- Flexible
- Widely supported
- Easy to use
- Easy to validate
- Fits the structure of UDC numbers

# The structure of composite UDC numbers

A composite UDC number can be represented as a **tree**.

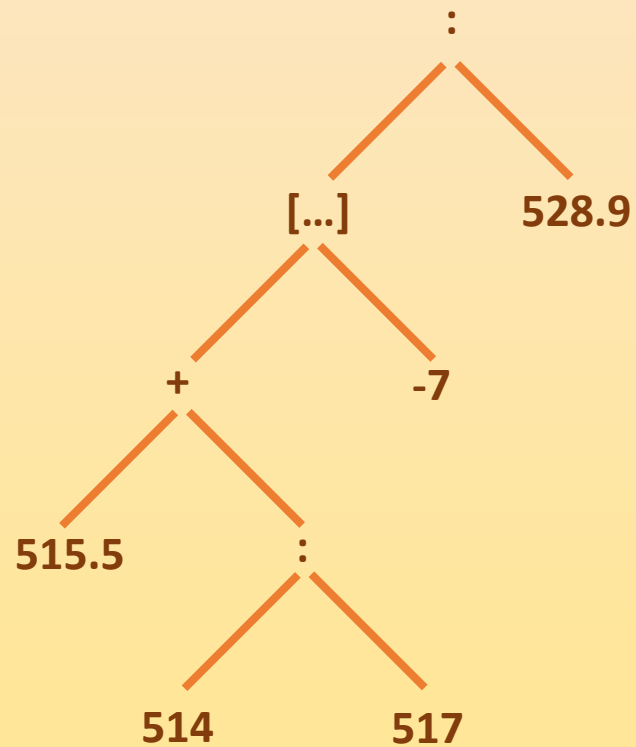
When there are multilevel relationships between its elements, their **priority order** is derived from the following conceptual definitions:

1. Simple numbers
2. Agglomeration of consecutive numbers. Subgrouping
3. Special and common facets
4. Phase relationships
5. Agglomeration



# The structure of composite UDC numbers

**[515.5+514:517]-7:528.9** Employing topological and analytic-geometrical methods in cartography



# The representation of UDC numbers

Each tree can be represented as an **XML**.

- The nodes of the tree are complex or single objects
- The possible object types are described by an **XML Schema Definition**

The XSD is **available at** <http://piros.udc-interpretor.hu#xsd>.

# The representation of UDC numbers

```
<ns:udc_concept xsi:schemaLocation="http://piros.udc-interpretter.hu/#xsd udc.xsd" xmlns:ns="http://piros.udc-interpretter.hu/#xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" udc_edition="2017" notation="515.1+514:517">
  <ns:description xml:lang="EN">Topology and analytical geometry</ns:description>
  <ns:main_concept>
    <ns:main_table_addition>
      <ns:main_concept order="1"><ns:main_table_number number1="515.1"/></ns:main_concept>
      <ns:main_table_relation order="2">
        <ns:main_concept order="1"><ns:main_table_number number1="514"/></ns:main_concept>
        <ns:main_concept order="2"><ns:main_table_number number1="517"/></ns:main_concept>
      </ns:main_table_relation>
    </ns:main_table_addition>
  </ns:main_concept>
</ns:udc_concept>
```

# Parsing UDC numbers by automatic means

- The automatic parsing of UDC numbers has been investigated since the **1960s** (cf. Rigby 1974)
- The most comprehensive research was conducted by **Gerhard Riesthuis** (1998)
- A similar research has been conducted in Hungary by **Gábor Mándy** (2013)

# Requirements for the interpreter

The next step is to implement an **interpreter** that converts UDC numbers **into the designed XML** format by

- retaining the **parts and the syntactic structure** of UDC numbers
- respecting all **UDC rules** by taking the **different editions** into account
- being fully **automatic** and **available online for** use by both humans and **programs**
  - <http://piros.udc-interpreter.hu#process>
- parsing the numbers in a **syntactic way**

# The functioning of the program

The interpreter is an **automaton** that recognizes the formal language determined by the classification system.

- **Input**

- The **UDC number**
- The year of the **UDC edition**

- **Output**

- The **XML representation** of the number or
- An **error message**

# Conversion to other formats

Processing XML may be too complicated for other applications, thus **conversion methods to other formats** have been implemented:

- **HTML** (display) format
- **KWOC** (Keyword Out of Context: list of the elements out of their context)
  - **HTML**
  - **JavaScript Object Serialization (JSON)** data-interchange format
    - Self-describing and easy to understand
    - Language-independent
    - Lightweight

# Conversion to other formats

```
{
  "concept": "378.4.007.1", "udc_edition": "1990",
  "pref_labels": {
    "pref_label_1": { "pref_label": "University senior officials. Chancellors, principals, rectors.",
      "language": "EN" }
  },
  "udc_numbers": {
    "number_1": { "notation": "378.4", "filing": "3T7T8C", "uri": "http://udcdata.info/025169", "pref_labels":
      { "pref_label_1": { "language": "EN", "pref_label": "Higher education. Universities. Academic study" } }},
    "number_2": { "notation": ".007.1", "filing": "P0T0T7T1C", "pref_labels": { "pref_label_1": {
      "language": "EN", "pref_label": "" } } }
  }
}
```



# A case study

The case study was conducted based on the open data of the **Digital National Library of Portugal** from The European Library (TEL) Open Dataset, which contains **13,741** unique UDC notations.

- Two **bugs** were found
- Five **unimplemented practices** to build numbers
- A number of typing **mistakes** and number building practices that didn't respect UDC rules by the indexers

# Changes in the XML

A new version (**version 2.1**) of the XML was released based on

- The **UDC Online English** (<http://udc-hub.com/> )
- The **case studies**
- Further research of the **literature**

It is **clearer**, better documented and **theoretically better established** than the previous version.

# Changes in the XML

The XML was **completed**, there is no further changes needed.

The most **important changes**

- The **special auxiliaries** were changed in order to handle even the most specific rules
- The **citation order** is stored for each element
- **Constraints** were corrected
- Handling special cases inside **auxiliaries of place** [place according to quadrants (*161/164*), measurements and dimensions (*18*)]
- Handling special cases inside **auxiliaries of language** (translations under =030.1/.9 and dialects in =...'276/'292).
- Handling the numbers of **main table numbers and independent common auxiliaries on the same level**

# The test set

- It contains more than **700 test cases**
- It can be used to **test any interpreter/parser**
- It will be **available** soon at <http://interpreter-eto.rhcloud.com/#test>

# The evaluation of the software

- **Aligning** the software **to the XML**
- Availability through a **RESTful** interface (Representational State Transfer)
  - An architecture style for distributed hypermedia applications
  - It would help other applications to apply the interpreter easier
- **Further machine-readable output formats**
  - **UNIMARC**
  - **RDF**

# Further research plans

Following the abovementioned improvements the **first phase** of the research will be completed.

The future research can focus on the **feasible applications** of the outputs of the previous phase.

# Further research plans

Analyzing the **experience**.

- The **test set** can be a basis for reviewing and analyzing the **current syntactic rules** of UDC
- The **further experience** gives us a chance of examining how the **revisions** performed in recent years **have helped in handling UDC notations**

# Further research plans

The **outputs of the interpreter** can serve as a basis for methods applying information regarding the syntactic structure of UDC numbers.

- **Quantitative studies** (cf. Smiraglia et al. 2013)
- **Similarity measurements** between composite subjects
- Developing **inference methods, searching and browsing** algorithms
- Building **permuted** and **KWIC-indexes**
- Implementing **intelligent classification interfaces**



# Thank you

**for your attention**

# References

Broughton, Vanda. 2015. *Essential Classification. Second Edition*. London: Facet Publishing.

Mándy Gábor. 2013. "A posztkoordináció esélyei az ETO-ban." *Könyvtári figyelő* 59: 65-84.

<http://ki.oszk.hu/kf/2013/04/a-posztkoordinacio-eselyei-az-eto-ban/>

Neelameghan, A. 1973. "Basic subjects". *Library Science with a Slant to Documentation* 10: 202-206.

Ogden, C. K., and Richards, I. A. 1946. *The meaning of meaning. a study of the influence of language upon thought and of the science of symbolism*. New York: Harcourt, Brace & World. Inc.

Piros Attila. 2017. "New automatic interpreter for complex UDC numbers." *Extensions and Corrections to the UDC*: 36-37 (2014-2015). [Forthcoming]

Ranganathan, S. R. 1967. *Prolegomena to library classification. 3rd ed*. New York: Asia Publishing House. <http://arizona.openrepository.com/arizona/handle/10150/106370>

Rigby, Malcolm. 1974. *Computers and the UDC. A decade of progress 1963-1973. FID 523*. The Hague: International Federation for Documentation.

Riesthuis, Gerhard J. A. 1998. "Zoeken met woorden: hergebruik van onderwerpsontsluiting." PhD diss., Amsterdam: Leerstoelgroep Boek-, Archief- en Informatiewetenschap.

Robinson, Geoffrey. 2003. "Citation Order in UDC." *Extensions and Corrections to the UDC* 25: 19-27.

Slavic, Aida. 2008. "Faceted classification: management and use." *Axiomathes* 18: 257-271. [doi:10.1007/s10516-007-9030-z](https://doi.org/10.1007/s10516-007-9030-z)

Smiraglia, Richard et al. 2013. "UDC in action." In *Classification and visualization: interfaces to knowledge: proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands*, edited by Aida Slavic, Almila Akdag Salah, and Sylvie Davies, 259-272. Würzburg: Ergon Verlag. <https://arxiv.org/abs/1306.3783>

# Acknowledgments

- my wife, Krisztina and my daughter, Eszter
- Dr. Aida Slavic
- my PhD supervisor Dr. István Boda
- Jonathan Wild
- Daniel Benediktsson