Multilingual UDC Summary Online Project: 2009 update

Aida Slavic
Associate Editor, UDC Consortium (United Kingdom)
Chris Overfield
Software developer (United Kingdom)
Gerhard Riesthuis
Associate Editor, UDC Consortium (The Netherlands)
Jiri Pika
ETH (Switzerland)

1. Project scope and objectives

UDC Summary (udcS) is a selection of around 2,000 UDC numbers intended for free use, training and research of the UDC, and is published as an online database at http://www.udcc.org/udcsummary/php/index.php.

This is the first time in the UDC's history that the scheme has been made available to any extent for free use in so many languages as a single service. By the end of 2009, this abridged scheme was available in 13 languages and at the time of writing this report there are already over 20 languages online. The UDC Summary is available in languages in which the UDC has never been translated before such as Armenian, Greek, and Hindi.

The objective behind the UDC Summary project was to illustrate the full power of the UDC, enhance its presence online and encourage the classification's use outside a traditional library setting. The udcS contains a selection of numbers that represent roughly 3.5% of the UDC Master Reference File. It contains main numbers, common auxiliary numbers and special auxiliary numbers and shows even coverage of all areas of knowledge. The data used in the UDC Summary mirror the structure of the complete UDC MRF database. This means that each class in the udcS is represented by a full set of data available for that class in the main system: verbal examples, scope notes, derivation instructions, application notes, examples of combination, see also references, search terms/class descriptors, alphabetical index entries, and mappings to other knowledge organization systems. In addition, each class number contains full administrative data (URI, date of introduction, last revision date, notation history etc.). Hence, udcS represents a fully functional UDC scheme suitable for online applications which can also be used as a UDC demonstrator for training, research, implementation testing and various information organization and retrieval purposes.

To all novice users and those unfamiliar with UDC, udcS can help by illustrating the structure of the UDC vocabulary and the basic principles of number building. But most importantly data exports and the planned data model, which will be available online for download by the end of 2010, will be of great help to those interested in the implementation of UDC in online searching and retrieval. Hence, although the udcS cannot, and is not intended to be used as an indexing tool, instead of the full UDC schedules, it can certainly help gain a better understanding of the scheme.

Most importantly, data exports in all languages will be available for free use and distribution under the Creative Commons Attribution Share Alike 3.0 license (CC-BY-SA). Hence the voluntary work put into translations feeds back directly into the community that needs it and can use these data.

2. The full picture

When completed, the UDC Summary, will contain the following elements:

1. UDC data:
- multilingual UDC Summary data in over 30 languages, updated and proofread by national editors/authorities
- subject-alphabetical indexes to the UDC Summary with around 10,000-14,000 language terms in as many languages as possible - with a controlled keyword index, relative index, and chain index
- UDC Summary mapping data to classification systems such as Dewey, Library of Congress Classification, Colon Classification and other national systems (e.g. German, Swedish, Chinese, Korean classification), thesauri and subject-heading systems
- exports: plain text exports with different levels of detail, machine readable exports (MARC, XML), machine-understandable exports, i.e. ontologies (SKOS, TopicMap)
- UDC Summary data available as linked (SKOS) data

2. Browsing & searching interface
3. Online collaborative maintenance/management tool
4. UDC associated instructional/training material
5. Free open source tools/software to use UDC data (parser, converter, validator)
6. Collecting user feedback/comments online

An important aspect of the UDC Summary will be its publication as linked data - which will enable udcS classes to be linked to and referenced by other programs, services and applications on the Web. The potential 'links' that we can envisage at this early stage are with wikipedia (dbpedia), geonames, and other knowledge organization systems or library catalogues, when published as linked data.

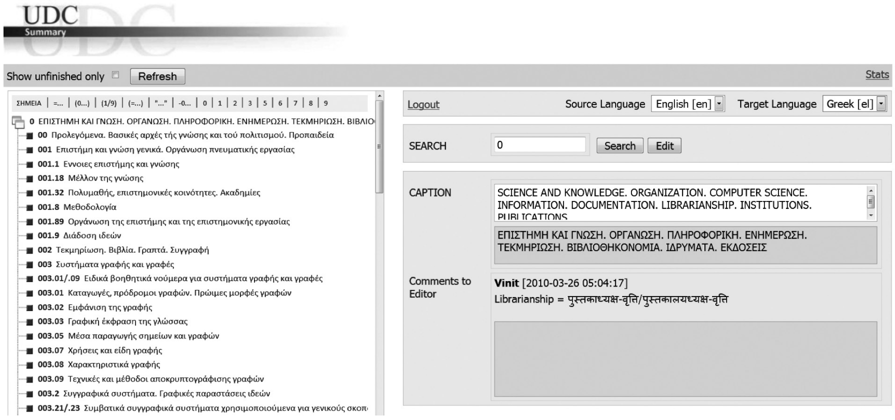
3. Management, infrastructure, procedure

udcS is held in a mySQL database that mirrors the structure of the main UDC MRF database and is aligned with the latest version of the UDC MRF. Thus each class in the UDC Summary exists in the UDC MRF and will be automatically updated with UDC MRF data. The database is held on the UDCC server.

The udcS project team consists of three content editors (Aida Slavic, Gerhard Riesthuis, Jiri Pika), a software developer (Chris Overfield) and over 60 translators. All project work is on a voluntary basis (see the list of collaborators at http://www.udcc.org/udcsummary/translation.htm). Each language has a main editor responsible for the quality of translation in a given language.

The work on each language starts by obtaining translated UDC data in the given language (if it exists) that can be automatically or semi-automatically aligned with and imported into the udcS database. If no data is available in digital form, volunteers are asked to translate the top two levels (around 100 classes) and the web interface. Upon the production of the top classes, the language data can be uploaded into the database and login access is provided to the online udcS Translation Tool for all translators assigned to this task. Normally, when data is obtained from existing editions there are still several hundred classes or notes that may need to be updated. The work on translation can continue either entirely online or using translation spreadsheets provided by the udcS editor. When spreadsheets with completed translations are returned they are uploaded and a new spreadsheet export produced.

Some translators prefer to work online and others like to work with spreadsheets or Word documents. Hence we are supporting both options and advise translators to choose whichever approach they find comfortable. If the two ways of working are combined we make sure that the process is coordinated. Frequently the online tool is considered important for final proofreading and corrections. Translators can leave comments in the field provided, either for the main udcS editor or for their colleagues. All comments left online are automatically posted via email upon saving the record (Figure 1).


UDC
Summary
Show unfinished only □
Refresh
Stats
ΣΗΜΕΙΑ | =... | (0...) | (1/9) | (=...) | "..." | -0... | 0 | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9
Logout
Search
Edit

0 ΕΠΙΣΤΗΜΗ ΚΑΙ ΓΝΩΣΗ. ΟΡΓΑΝΩΣΗ. ΠΛΗΡΟΦΟΡΙΚΗ. ΕΝΗΜΕΡΩΣΗ. ΤΕΚΜΗΡΙΩΣΗ. ΒΙΒΛΙΟ

- **00** Προλεγόμενα. Βασικές αρχές τής γνώσης και τού πολιτισμού. Προπαιδεία
- **001** Επιστήμη και γνώση γενικά. Οργάνωση πνευματικής εργασίας
- **001.1** Εννοιες επιστήμης και γνώσης
- **001.18** Μέλλον της γνώσης
- **001.32** Πολυμαθής, επιστημονικές κοινότητες. Ακαδημίες
- **001.8** Μεθοδολογία
- **001.89** Οργάνωση της επιστήμης και της επιστημονικής εργασίας
- **001.9** Διάδοση ιδεών
- **002** Τεκμηρίωση. Βιβλία. Γραπτά. Συγγραφή
- **003** Συστήματα γραφής και γραφές
- **003.01/.09** Ειδικά βοηθητικά νούμερα για συστήματα γραφής και γραφές
- **003.01** Καταγωγές, πρόδρομοι γραφών. Πρώιμες μορφές γραφών
- **003.02** Εμφάνιση της γραφής
- **003.03** Γραφική έκφραση της γλώσσας
- **003.05** Μέσα παραγωγής σημείων και γραφών
- **003.07** Χρήσεις και είδη γραφής
- **003.08** Χαρακτηριστικά γραφής
- **003.09** Τεχνικές και μέθοδοι αποκρυπτογράφισης γραφών
- **003.2** Συγγραφικά συστήματα. Γραφικές παραστάσεις ιδεών
- **003.21/.23** Συμβατικά συγγραφικά συστήματα χρησιμοποιούμενα για γενικούς σκοπ

Source Language English [en]
Target Language Greek [el]

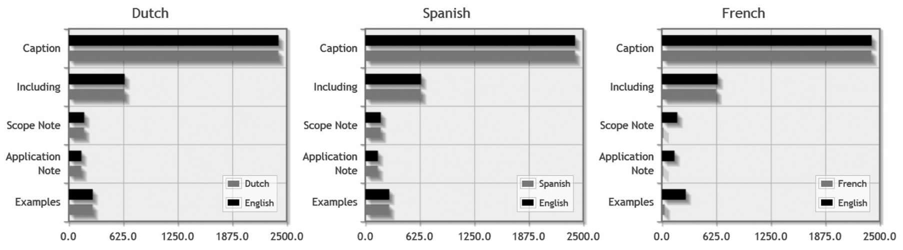
SEARCH 0

CAPTION
SCIENCE AND KNOWLEDGE. ORGANIZATION. COMPUTER SCIENCE. INFORMATION. DOCUMENTATION. LIBRARIANSHIP. INSTITUTIONS. PUBLICATIONS
ΕΠΙΣΤΗΜΗ ΚΑΙ ΓΝΩΣΗ. ΟΡΓΑΝΩΣΗ. ΠΛΗΡΟΦΟΡΙΚΗ. ΕΝΗΜΕΡΩΣΗ. ΤΕΚΜΗΡΙΩΣΗ. ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑ. ΙΔΡΥΜΑΤΑ. ΕΚΔΟΣΕΙΣ

Comments to Editor
Vinit [2010-03-26 05:04:17]
Librarianship = पुस्तकाध्यक्ष-वृत्ति/पुस्तकालयध्यक्ष-वृत्ति

Figure 1 – **udcS translation tool (access to Greek translation)**

Translation progress can be observed at the udcS Completion Statistics page available at http://www.udcc.org/udcsummary/stats/php/trans_stats.php. This shows the percentage of fields translated for each language. On the graphic display the top bar (appears in orange on the web) shows the English source data and the second (turquoise) bar the status of a given language. The statistics displayed online are 'live', i.e. the view changes with every record translated and saved in any language (Figure 2).



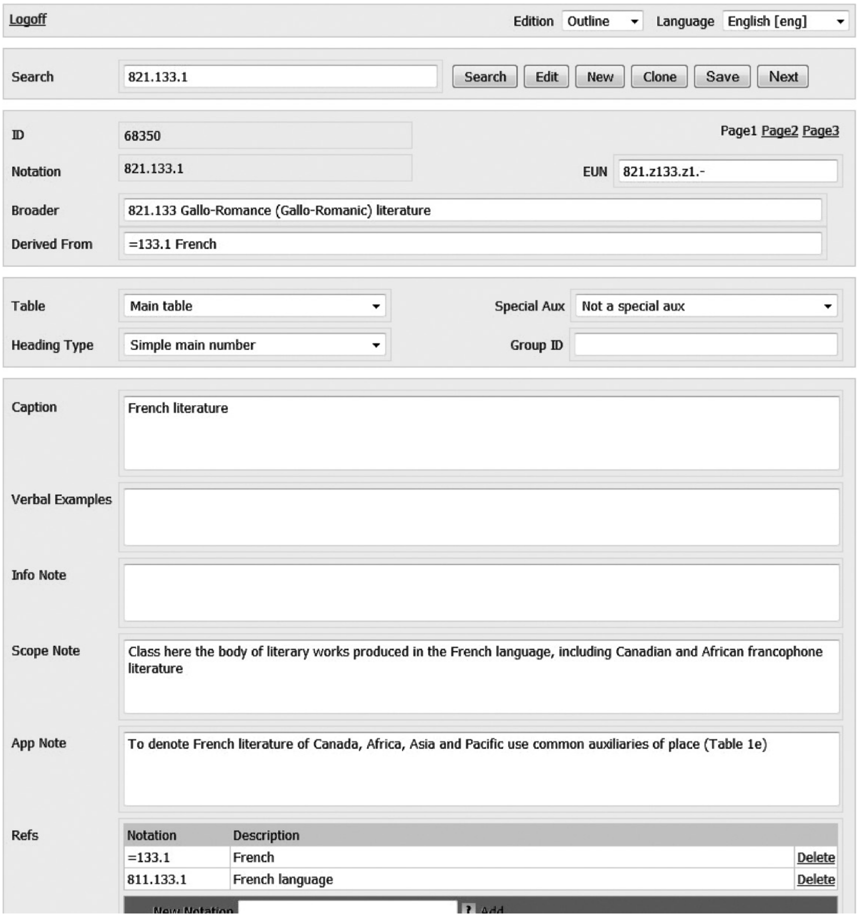
Field	% completion
Caption	100.0
Including	99.8
Scope Note	100.0
Application Note	100.0
Examples	100.0

Field	% completion
Caption	100.0
Including	99.8
Scope Note	100.0
Application Note	100.0
Examples	100.0

Field	% completion
Caption	100.0
Including	97.7
Scope Note	10.2
Application Note	10.5
Examples	12.1

Figure 2 – **Translation progress statistics per language**

When a translator logs in to the online translation tool, they are automatically given access only to the language they are working on. However, for central maintenance there is a main UDC Summary editor online which provides access to the complete database content and is maintained by the udcS editor-in-chief. The main editor allows access to all languages and all fields of the database, and permits the entry of new records (Figure 3).



Logoff   Edition: Outline   Language: English [eng]

Search: 821.133.1   Search | Edit | New | Clone | Save | Next

Page1 Page2 Page3

ID: 68350

Notation: 821.133.1   EUN: 821.z133.z1.-

Broader: 821.133 Gallo-Romance (Gallo-Romanic) literature

Derived From: =133.1 French

Table: Main table   Special Aux: Not a special aux

Heading Type: Simple main number   Group ID:

Caption: French literature

Verbal Examples:

Info Note:

Scope Note: Class here the body of literary works produced in the French language, including Canadian and African francophone literature

App Note: To denote French literature of Canada, Africa, Asia and Pacific use common auxiliaries of place (Table 1e)

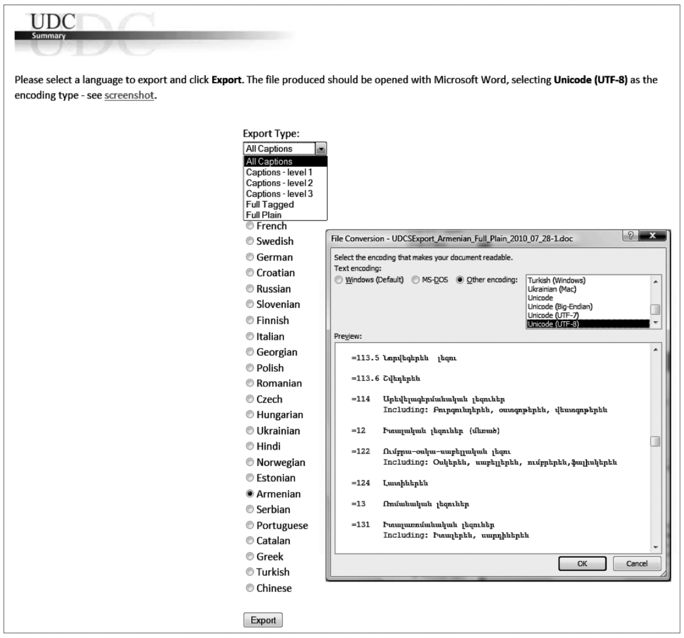
Refs

Notation	Description	
=133.1	French	Delete
811.133.1	French language	Delete

New Notation: ? Add

Figure 3 – **The first page of the main udcS editor**

One of the most important tasks is the proofreading and checking of the schedules upon completion or during the process of translation. To enable off-line proofreading and checking at this stage we are using a data export service online which produces exports of several types and levels of detail directly from the database. The export service will be expanded later to provide more complex exports (Figure 4).



UDC Summary

Please select a language to export and click **Export**. The file produced should be opened with Microsoft Word, selecting **Unicode (UTF-8)** as the encoding type - see screenshot.

Export Type: All Captions

○ French
○ Swedish
○ German
○ Croatian
○ Russian
○ Slovenian
○ Finnish
○ Italian
○ Georgian
○ Polish
○ Romanian
○ Czech
○ Hungarian
○ Ukrainian
○ Hindi
○ Norwegian
○ Estonian
● Armenian
○ Serbian
○ Portuguese
○ Catalan
○ Greek
○ Turkish
○ Chinese

File Conversion - UDCSExport_Armenian_Full_Plain_2010_07_28-1.doc

Select the encoding that makes your document readable.

Text encoding:

○ Windows (Default) ○ MS-DOS ● Other encoding: Unicode (UTF-8)

Preview:

```
=113.5 Նորվեգերեն  լեզու

=113.6 Շվեդերեն

=114   Արեւելագերմանական լեզուներ
       Including: Բուրգունդերեն, օստգոթերեն, վեստգոթերեն

=12    Իտալական լեզուներ (մեռած)

=122   Ումբրա-օսկա-սաբելյական լեզու
       Including: Օսկերեն, սաբելլերեն, ումբրերեն,ֆալիսկերեն

=124   Լատիներեն

=13    Ռոմանական լեզուներ

=131   Իտալառոմանական լեզուներ
       Including: Իտալերեն, սարդիներեն
```

Figure 4 – **udcS export tool online**

4. Concluding remarks

The advantages of having the UDC Summary available online are self-evident and easy to understand. Further to this, work on this project has yielded benefits to the UDC editorial and software development team. In the short period we have worked on the udcS, we have managed to engage and collaborate with a greater number of colleagues worldwide including UDC editors of national editions, lecturers, librarians and researchers. The majority of the UDC Editorial team and the UDC Advisory Board members, including the UDCC office are all actively involved either by doing translations or by helping establish contacts with other translators and publishers. In the process of the udcS preparation and translation we collected feedback, comments and suggestions for improving the schedules and were able to start implementing some of these suggestions in the UDC MRF at the end of 2009.

The project enabled us to validate UDC MRF data and test the management of multilingual data online. In the process we explored data conversion services and tested various interface designs that the UDC Consortium may consider implementing in the future as full data services. We have developed methods and software and we have measured the effort required in aligning old editions with the UDC MRF, gaining better understanding of the procedures, time involved and potential costs. It is evident that if such a service were offered to UDC publishers it would be of great benefit as it would shorten the production time of new editions.

On a more technical level, working on multilingual UDC data helped in testing the performance of the new database, selecting the most appropriate character coding standards, etc. All these results contribute directly to the improvement of the UDC MRF and the services based on it. Throughout 2010 the udcS will be in full development as we will be adding functionalities and importing more language data. By the end of the year we hope to have over 30 languages translated and the majority of search and export services fully functional.