Assembly of Gemet 2001

Report on the assembly of Gemet 2001 including the integration into THESshow

September 2001
Hermann Stallbaumer

Preparing the database

The creation of the Gemet 2001 database started from the Gemet 2000 database from September 2000.

The following steps were performed import the the additional language data:

- Rearangement of languages
- Adding terms for Baskque, Bulgarian, Russian and Slovenian
- Adding improvements for Portuguese and Swedish
- Increasing the possible number of languages from 15 to 20 in Thesshow

Language arrangement in Gemet 2001

LangPtr	Language	Font	Order	Abbrev	Codepage	Charset
1	English	Times New Roman	1	eng	0	Iso-Latin 1
2	Danish	Times New Roman	2	dan	0	Iso-Latin 1
3	Dutch	Times New Roman	3	Dut	0	Iso-Latin 1
4	German	Times New Roman	4	Ger	0	Iso-Latin 1
5	Norwegian	Times New Roman	5	Nor	0	Iso-Latin 1
6	Swedish	Times New Roman	6	Swe	0	Iso-Latin 1
7	French	Times New Roman	7	Fre	0	Iso-Latin 1
8	Italian	Times New Roman	8	Ita	0	Iso-Latin 1
9	Portuguese	Times New Roman	9	Por	0	Iso-Latin 1
10	Spanish	Times New Roman	10	Spa	0	Iso-Latin 1
11	Greek	Times New Roman	11	Grk	161	Iso-Latin 7
12	Finnish	Times New Roman	12	Fin	0	Iso-Latin 1
13	Hungarian	Times New Roman	13	Hun	238	Iso-Latin 2
14	Bulgarian	Times New Roman	14	bul	204	Iso-Latin 5
15	Russian	Times New Roman	15	rus	204	Iso-Latin 5
16	Slovak	Times New Roman	14	slo	238	Iso-Latin 2
17	Slovenian	Times New Roman	17	slv	238	Iso-Latin 2
18	Basque	Times New Roman	18	baq	0	Iso-Latin 1
19	US-English	Times New Roman	19	Usa	0	Iso-Latin 1

Rearrangement of languages

In Gemet 2.0 languages were arranged in two groups, germanic languages and roman languages. Within the groups languages were ordered alphabetically. For English is the leading language, the Germanic group was ordered first, the roman group behind.
In Gemet 2001 this kind of ordering no longer is possible for there are languages added which do not belong to either of the two groups.

In Gemet 2001 the languages are divided into the following groups:

Germanic languages:

English
Danish
Dutch
German
Norwegian
Swedish

Romance languages:
French
Italian
Portuguese
Spanish

Greek language (A language family by its own)
Greek

Finno-Ugric languages (The only non-indogermanic language family in Europe)
Finnish
Hungarian

Slawic languages:
Bulgarian
Russian
Slovak
Slovenian

Basque language (An isolated language belonging to no language family, the only living pre-indogermanic language in europe)
Basque

US-English was put at the end as it is the only non European country present in Gemet.

Sources:
www.weikopf.de (very informative on this subject)
www.buber.net (Basque related)
www.ethologue.com

Language abbreviations :

Language abbreviations are due to ISO-Standard 639-2 alpha-3.
The standard now recommends, in cases of two abbreviations, the use of the bibliographic code. This lead to the rearrangement of Dutch/German

Import of Languages

Swedish

Was already present in Gemet 2000. Some changes were requested.

15 terms to change were supplied. The corresponding entries are changed.
5 new Synonyms added

English term	Correct Swedish term	Synonym 1	Synonym 2
vermin	ohyra	Skadedjur	
abandoned vehicle	övergivet fordon		
weed control	ogräsbekämpning		
International Monetary Fund	Internationella valutafonden		
catchment area	avrinningsområde	tillrinningsområde	
literature evaluation	litteraturutvärdering	litteraturstudie	
agricultural building	ekonomibyggnad		
urban sprawl	tätortsutbredning		
legal procedure	rättsligt förfarande		
green building	ekohus		
bush clearing	buskröjning	Slyröjning	röjning av sly
climatic zone	klimatzon		
employment level effect	sysselsättningseffekt		
form of government	styresform		
agricultural real estate	jordbruksfastighet		

4868 Descriptors and 79 Nondescriptors are in the Swedish data content

427 entries are still missing.

120 Descriptors are duplicate entries

Portuguese

Was also already present in Gemet 2000.

5295 changed Descriptor entries were delivered and imported without problems.

The following table of Nondescriptors were requested to be removed. Instead of referring to Portuguese Nondecriptor ID's, by chance corresponding English Nondescriptor ID's are given. For Nondescriptors and their ID's in principal are not correlated between languages referencing via foreign language ID's is very likely to be wrong. Referencing into Portuguese Nondescriptor content was therefore performed manually. The result is given in the following table.

DescNr	NonDescr old	English	NonDescriptor New
12100	lençóis aluviais	alluvial sheet	baixas aluviais
14834		system analysis	
400	dispositivos de análise	analytical device	instrumentos de análise
4237	pousio industrial	industrial fallow	terrenos industriais abandonados
4585	horários de trabalho	labour scheduling	planeamento do trabalho
4674	registo de propriedades (rural)	land title register	registo de títulos de propriedade
4703	lavandarias	laundry service	tratamento de roupa
4869	lixiviador	leachate	lixiviado
5075	terra pantanosa	marshland	terras pantanosas
6843	processo de depuração	purification process	processos de depuração
7196	embalagens com retorno	returnable packaging	recipientes com retorno
7199	embalagens reutilizáveis	reusable packaging	recipientes reutilizáveis
7332	lixeiras	rubbish dump	lixeiras
7635	estação de tratamento de efluentes	sewage treatment plant	estações de tratamento de efluentes
8177	águas subterrâneas	subterranean water	águas intersticiais
8863		used oil	
9115	minimização de resíduos (total)	waste minimising	diminuição de resíduos

32 Nondescriptors were requested to be added. Again the English Nondescriptor references were given. The terms were added with the following sql statements. After introducing artificial Nondescriptor pointers and adding 1000000 to each DescPtr field the according Descriptor ID's were added by:

```
UPDATE DISTINCTROW PORTSYNA INNER JOIN ConNonDesc1 ON PORTSYNA.NDescPtr =
ConNonDesc1.NonDescNr SET PORTSYNA.DescPtr = [ConNonDesc1].[DescPtr];
```

The Portuguese Connectivity table was updated by:

```
INSERT INTO ConNonDesc9 ( NonDescNr, DescPtr )
SELECT DISTINCTROW PORTSYNA.ID, PORTSYNA.DescPtr
FROM PORTSYNA;
```

5295 Descriptors and 146 Nondescriptors are in the Portuguese data content

No duplicates neither in the Descriptor- nor in the Nondescriptortable were found.

Group and Theme entries were updated according to the supplied material.

Bulgarian

5273 Descriptors were delivered from which 5268 could be referenced automatically. For no Term ID's were available the data was referenced via the English entries.

The following entries were handled manually:

Canid (canine)	(1133)
Lome convention	(1138)
dust emission	*unable to reference*
Morphology (biol.)	(5369)
overflow outlet	(13570)

13 of these entries were left untranslated so actually 5259 Descriptors are available

65 Descriptors are duplicates

39 Groups and 40 Themes were imported

Not imported were the terms from Group NAT which is ommitted in Gemet-2000.

Russian

5295 out of 5311 Descriptors were imported.

130 Descriptors are duplicates

39 Groups and 40 Themes were imported

Two additional tables were delivered for use as Nondescriptors:

Unfortunately all ID's are not from Gemet. So referencing was performed by English entries. For these entries are handwritten some entries could not be referenced automatically. See the following table for further information.

Code	Descriptor	Related term	Manual change	ID
29	action	акция		70
57	advice	консультация		130
751	contamination	заражение		1752
788	cost	цена, расходы		1824
861	decomposition	распад		2014
870	defence	оборона		2031
1120	elasticity	упругость		2582
1437	feed-in current	ток питания	Not in Gemet	
1718	grinding	дробление		3774
1804	holiday	каникулы, выходной день		3986
1811	household	домочадцы		4015
1987	in situ	"в месте нахождения"		4359
2090	land	суша, местность		4599
2123	landscape management	уход за ландшафтом		4658
2130	lanslide	обвал	landslide	4668
2171	levy	налог		4784
2245	luminocity	яркость	luminosity	4397

3198	refrigeration	замораживание		7044
3220	ordinance	закон, статут		7087
3399	seal	затвор	seal (technical)	7519
3504	sluice	затвор		7742
3697	storm	шторм		8115
3714	submarine	подводный		8159
4004	value	ценность, стоимость		8890
4062	wastage	усушка, изнашивание		9040
4533	calculation	подсчет		11113
4657	litigation	тяжба, спор		11563
5134	allowance	поправка, скидка, норма отпуска, разрешение, фора		13560

Code	Descriptor	Synonym	zz
493	carbon dioxide	двуокись углерода, углекислый газ	1168
495	carbon monoxide	окись углерода, угарный газ	1173
534	cephalopod	головоногие	1265
1349	enzyme	фермент	2945
1988	inspection	проверка	4360
2722	overburden	покрывающая порода	5956
3062	protection	защита	6748
3182	recycling	рециклинг	7015
3199	refrigerator	рефрижератор	7045
3202	refuse collection vehicle	мусоровоз	7050
4326	geographic circque	кар	10146
4412	local afforestation	лесонасаждение	10683
4705	sulphur monoxide	окись серы	11734
4790	bog	топь, болото	12161
4881	vitrification	переход в стеклообразное состояние	12611

From these tables 42 Nondescriptors were generated including one duplicate.

The assignment was performed by sql:

```
UPDATE DISTINCTROW [Desc] INNER JOIN Rus_Syn ON Desc.Desc1 =
Rus_Syn.Descriptor SET Rus_Syn.zz = [Desc].[DescNr];
```

The Nondescriptor entries were created by:

```
INSERT INTO NonDescr15 ( NonDesc1, NonDescNr )
SELECT DISTINCTROW Rus_Rel.[Related Term], Rus_Rel.NonDescNr
FROM Rus_Rel;
```

The Nondescriptor connectivity was created by:

```
INSERT INTO ConNonDesc15 ( DescPtr, NonDescNr )
SELECT DISTINCTROW Rus_Rel.DescNr, Rus_Rel.NonDescNr
FROM Rus_Rel;
```

Slowenian

5295 Descriptors out of 5311 delivered were imported. Not imported were the terms from Group NAT which is ommitted in Gemet-2000.
All Groups and Themes were supplied and imported. No Nondescriptors were delivered.
157 Descriptors are duplicates.
Remark: A few (9) entries contain more than one Descriptor separated by a semicolon. This technique is not state of the art and makes the use of the data content almost impossible for indexing purposes.

Baskque

5295 Descriptors out of 5311 delivered were imported. Not imported were the terms from Group NAT which is ommitted in Gemet-2000.
All Groups and Themes were supplied and imported. No Nondescriptors were delivered.
101 Descriptors are duplicates.
Remark: Almost 2000 entries contain more than one Descriptor separated by a semicolon. This technique is not state of the art and makes the use of the data content almost impossible for indexing purposes.

Gemet 2001 Summary

Nr	Language	Language Code	Charset	Nondescriptors	Entries missing
1	English	ISO 8859-1	0	1537	
2	Danish	ISO 8859-1	0	1283	18
3	Dutch	ISO 8859-1	0	1939	
4	German	ISO 8859-1	0	534	
5	Norwegian	ISO 8859-1	0	82	
6	Swedish	ISO 8859-1	0	79	427
7	French	ISO 8859-1	0	289	
8	Italian	ISO 8859-1	0	682	
9	Portuguese	ISO 8859-1	0	146	
10	Spanish	ISO 8859-1	0	212	
11	Greek	ISO 8859-7	161	149	4
12	Hungarian	ISO 8859-2	238		1
13	Finnish	ISO 8859-1	0	80	3
14	Bulgarian	ISO 8859-5	204		36
15	Russian	ISO 8859-5	204	42	
16	Slovak	ISO 8859-2	238		67
17	Slowenian	ISO 8859-2	238		
18	Basque	ISO 8859-1	0		
19	US-English	ISO 8859-1	0	1260	