

## Open source information for export control

Cristina Versino, Camelia Ignat and Louis-Victor Brill

European Commission, Joint Research Centre, via Fermi, 21020 Ispra, Italy  
e-mail: cristina.versino@jrc.it, camelia.ignat@jrc.it, louis-victor.brill@jrc.it

### **Abstract:**

*This paper presents some preliminary work in the area of export control for nuclear proliferation. The long term goal is to access and analyse data that may relate to the trade of nuclear dual-use items and technologies: this data is available through various open sources. The immediate goal is to work on a table of correspondence between codes that describe, on one side, nuclear items and the nuclear- dual-use, and, on the other side, goods as they are declared in trade databases. We present a first exploration on the use of language technology to help bridge these two, very different, worlds of terms.*

**Keywords:** export control, dual-use, combined nomenclature, language technology, open source.

### **1. Introduction**

Non-proliferation agreements are inter alia reflected under export control regimes: these pose restrictions on the trade of items which can assist the manufacturing of chemical, biological or nuclear weapons or other nuclear explosive devices as well as missile technology.

In the context of nuclear non-proliferation, the **Nuclear Suppliers Group** (NSG) authored specific guidelines describing items of interest to nuclear proliferation and aiming at controlling the export of nuclear material, equipment and technology (INFCIRC/254/Part 1) [1] as well as for the transfer of nuclear-related dual-use equipment, materials, software and related technology (INFCIRC/254/Part 2) [2].

At European level, the NSG guidelines are incorporated as part of the broader **Council Regulation 1334/2000** [3] and amendments [4]. This single regulation implements four internationally agreed dual-use controls, namely the Wassenaar Arrangement, the Missile Technology Control Regime (MTCR), the Australia Group the Chemical Weapons Convention (CWC) and the Nuclear Suppliers Group. The regulation identifies a single list of items to be controlled over the four Regimes: items are enumerated by a coding scheme, here referred to as '**EU dual-use codes**' (**EU-DU-C**), whose structure and granularity is driven by proliferation concerns.

Databases on international trade present an interest in export control. At European level, COMEXT [5] by EUROSTAT collects data on trade between EU Member States and non-member countries. Another example is COMTRADE [6]: maintained by the Statistics Division of the United Nations, it provides a worldwide view on trade. **In these databases trade data is reported according to product classification schemes that are independent from EU-DU-C.** For instance, in COMEXT items are indexed by '**Combined Nomenclature codes**' (**CN-C**) [7] whose structure and granularity reflect customs tariffs and not proliferation concerns.

**As a result, to access trade data relevant to nuclear non-proliferation, it is first necessary to establish a mapping between the various sets of codes, e. g. EU-DU-C to CN-C.**

The paper reports on exploratory work carried out on the official EU '**Correlation Table**' that maps **EU-DU-C to CN-C, with a focus on items identified in the NSG guidelines.**

The first goal of the exercise is to evaluate and eventually improve the quality of this correspondence table by coupling nuclear domain expertise with language technology tools.

The longer term objective is to use the table as a key to access trade data for verification; in doing so, we are aware that any mapping between coding schemes inevitably introduces approximate queries, but these approximations cannot be avoided. Exporters of dual-use items and custom officers are both

CN-C	EU-DU-C
84011000	0A001a
84014000	0A001b
84261100	0A001c
84261900	0A001c
...	...

**Table 1:** First lines the EU Correlation Table.

confronted with the same dilemma: traders are required to attribute CN codes to their exports (EU-DU-C to CN-C); vice versa, customs officers need to identify the nature of exports from CN codes (CN-C to EU-DU-C). They both use the Correlation Table, each in the appropriate direction.

## 2. The official EU Correlation Table

At EC level, DG Taxation and Customs Union (TAXUD) maintains the official EU Correlation Table that creates a correspondence between combined nomenclature codes (CN-C) and EU dual-use codes (EU-DU-C). Updates of the table reflect amendments of the Council Regulation on dual-use [3, 4] as well as the yearly revision of the Combined Nomenclature [7].

The mapping provided by the table is a not a one-to-one mapping: the same CN-C may describe more than one EU-DU-C; likewise, several CN-C may be associated to the same EU-DU-C.

Few entries of the Correlation Table are shown in Table 1. The first line, for instance, associates:

*Nuclear reactors [Euratom] (84011000)*

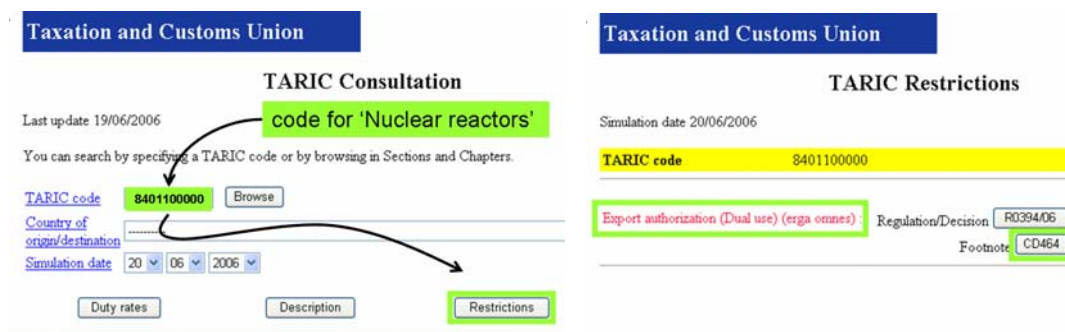
and

*Nuclear reactors capable of operation so as to maintain a controlled self-sustaining fission chain reaction (0A001a).*

To date, the Correlation Table serves the practical purpose of informing exporters and custom officers of Member States on the restrictions that apply to the trade of goods defined in the Council Regulation on dual-use [3, 4].

Technically, the Correlation Table is part of the 'Integrated tariff of the European Communities' (TARIC [8]). TARIC incorporates the Community legislation on trade concerning tariff suspensions, quotas, import/export prohibitions, surveillance, restrictions, etc. It identifies goods by TARIC codes, a subdivision of CN-C codes that adds 2 rightmost digits.

TARIC is used by the Commission and the Member States for the purpose of applying Community measures relating to imports and exports. A web site [9] is dedicated to the consultation of TARIC (Figure 1, left). For example, by entering the TARIC code 8401100000 (corresponding to 'Nuclear reactors') an exporter is made aware of the restrictions that apply to the trade of this category of goods (Figure 1, right): namely, that an export authorization is required because of Regulation R0394/06. The footnote CD464 also gives access to EU-DU-C corresponding to 8401100000 (i.e., 0A001a, not



**Figure 1:** The TARIC consultation site provides information about trade restrictions.

Nr	Dual-use Code	CN Code	Meaning of CN Code
1	0A001a	8401_10_00	Nuclear reactors [Euratom] .
2	0A001b	8401_40_00	Parts of nuclear reactors, n.e.s. [Euratom] .
3	0A001c	8426_11_00	Overhead travelling cranes on fixed support .
		8426_19_00	Overhead travelling cranes, transporter cranes, gantry cranes, bridge cranes and mobile lifting frames (excl. overhead travelling cranes on fixed support, mobile lifting frames on tyres, straddle carriers and portal or pedestal jib cranes) .
		8426_99_00	Ships" derricks; cranes, incl. cable cranes (excl. overhead travelling cranes, transporter cranes, gantry cranes, portal or pedestal jib cranes, bridge cranes, mobile lifting frames and straddle carriers, tower cranes, works trucks fitted with a crane, mobile cranes and cranes designed for mounting on road vehicles) .
		8428_90_97	Lifting, handling, loading or unloading machinery, n.e.s. .
4	0A001d	8401_40_00	Parts of nuclear reactors, n.e.s. [Euratom] .

Figure 2: Extract from the Nuclear Correlation Table with the meaning of CN codes added.

shown in Figure 1).

### 3. Focusing on the Nuclear Correlation Table

The Correlation Table covers the whole Council Regulation on dual-use which is broader than the nuclear focus offered by the NSG guidelines.

Related to this, it is to be noted a *different use of terms within the nuclear community and between those who refer to the Council Regulation on dual-use*. For the nuclear community, NSG Part1 is purely nuclear (i.e., it includes “nuclear” material, equipment designed for nuclear industry and declared as such) and it is also referred to as ‘trigger list’<sup>1</sup>. NSG Part2 covers those equipment that can be used inter alia for nuclear explosive activity, hence the qualification of ‘dual-use’ equipment. On the other hand, the Council Regulation does not make this distinction and calls dual-use items appearing both in NSG Part1 and Part2. Even more, all items listed in the Regulation are called dual-use and they derive from four export control regimes that originated independently and intersect on some items.

As a consequence, if one is primarily interested in trade data related to NSG Part 1 and Part 2, three steps need to be accomplished:

1. Tag items of interest in the NSG guidelines using the native NSG-C coding system;
2. Retrieve these items within the Council Regulation on dual-use, i.e. establish the mapping NSG-C → EU-DU-C.
3. Use this mapping to identify within the EU Correlation Table all and only the lines that relate to NSG-C: these lines together make the **Nuclear Correlation Table**. This is equivalent to derive the correspondence NSG → CN-C.

Figure 2 shows part of the Nuclear Correlation Table sorted by EU-DU-C. The meaning of each CN code (derived from the ‘Structure and self-explanatory texts’ of the Combined Nomenclature downloadable from [7]) has been added for clarity.

Table 2 provides numbers on the size of the Nuclear Correlation Table in comparison to the complete Correlation Table.

Correlation Table	Number of rows	Distinct EU-DU-C	Distinct CN-C
Complete	3189	548	928
Nuclear	571	116	372

Table 2: Comparing the size of the complete and of the nuclear Correlation Table.

<sup>1</sup> One should note that import / export of “nuclear” materials are addressed by the Euratom Treaty and derived regulation.

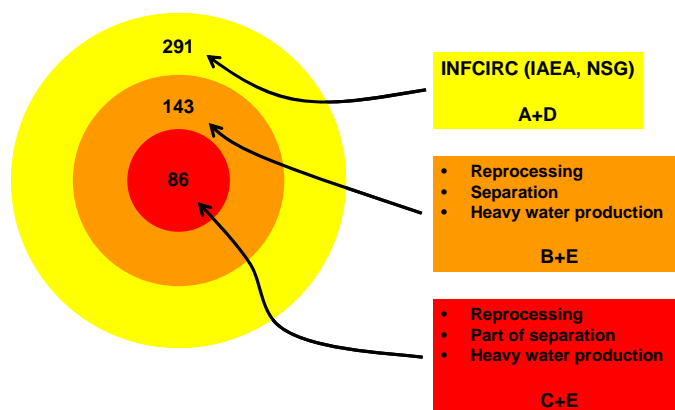


Figure 3: Priority levels applied to items in the NSG guidelines.

#### 4. The 'size' of the nuclear export control problem

To have an idea of the 'size' of the nuclear and nuclear-related export control problem as depicted by the NSG guidelines, we counted the items listed in these documents.

- The total number of items amounts to 291.
- Of these, 133 come from NSG Part 1 [1].
- 158 come from NSG Part 2 [2].

A priori, all NSG items are relevant for the nuclear export control problem. Nevertheless, to manage the size of the problem, we can identify **different levels of attention on items** pertaining to processes of the fuel cycle that are more critical or appropriate with respect to nuclear proliferation. Indeed, regarding enrichment, not all isotope separation methods are equally interesting from the proliferation point of view.

The following **prioritization of NSG items** is proposed.

Three levels of priority A, B and C are defined on NSG Part1.

- Level **A** includes all items → **133 items**.
- Level **B** includes 'Equipment and Non-nuclear Materials' items that fall in sections 3. (Reprocessing) and 5. (Separation of isotopes of uranium) → **98 items**.
- Level **C** includes 'Equipment and Non-nuclear Materials' items that fall in sections 3. (Reprocessing) and sections 5.1., 5.2. and 5.7. → **41 items**.

Two levels of priority D and E are defined on INFCIRC/254/Part 2.

- Level **D** includes all items → **158 items**.
- Level **E** includes items that fall in sections 3. (Uranium isotope separation equipment and components) and 4. (Heavy water production plant related equipment) → **45 items**.

For illustration purposes, Figure 3 gives a view on how these prioritized items add in numbers by combining:

- A and D → **291 (yellow)**
- B and E → **193 (orange)**
- C and E → **86 (red)**

'Yellow' covers the complete NSG guidelines, Part 1 and Part 2. 'Red' are the core items one may want to control with the highest level of attention. 'Orange' is an intermediate level.

#### 5. Language technology and the Nuclear Correlation Table

There are two motivations for implementing language technology in relation to the Nuclear Correlation Table.

1	<b>TERM</b>	<b>LH_MEASURE</b>	15	centrifuge	105,1636299	29	vapour	62,38738146
2	resistant	313,5056238	16	equipment	90,71433366	30	suitable	56,78168827
3	uranium	312,4852022	17	tantalum	87,59108827	31	collector	56,71265678
4	corrosion	279,0118061	18	fluorocarbon	87,59108827	32	frequency	56,24154641
5	separation	275,6854046	19	design	86,03753292	33	exchanger	54,32328568
6	protect	135,3947083	20	laser	85,57661264	34	concentrated	54,11391155
7	specially	132,9561461	21	rotor	84,92623772	35	magnet	51,3371621
8	diameter	128,5308539	22	component	73,72527024	36	crucible	50,916371
9	graphite	116,2967447	23	nozzle	70,74382037	37	plant	50,88764534
10	ion	111,5755067	24	cryogenic	70,01370589	38	compressor	50,67530203
11	gas	110,7547625	25	hydrochloric	69,38046152	39	gaseous	49,56832132
12	tube	109,9768268	26	therefor	65,41115518	40	temperature	49,55252851
13	isotope	107,4537651	27	ion-exchange	63,96860447	41	cylindrical	46,05670189
14	capable	106,6827657	28	consist	63,45344924	42	heat	44,28932461

**Figure 4:** Ranking of single words by their significance for the paragraph 0B001 of the Council Regulation.

First, before using the Table to access trade databases, we want to be in a position to assess the quality of the mapping between EU-DU-C and CN-C. The issue is to establish to what extent this mapping is accurate and whether certain parts of it need and can be improved. In general, answering these questions requires deep knowledge of the Combined Nomenclature. Nevertheless, since a textual description is attached to both EU-DU-C and CN-C, we use language technology to match these descriptions in an automatic way, i.e. without the help of experts of the Combined Nomenclature.

Second, considering that both the Council Regulation and the CN exist in all 23 official languages of the EU, using language technology opens the possibility to perform this evaluation in the languages of preference of the users, possibly mixing languages in a team.

The idea underlying the language technology approach followed here is to suggest associations between EU-DU-C and CN-C by focusing on items referred by the same **words of significance** (see paragraph 3.1 in [10]) **both in the EU-DU-C and CN-C textual descriptions**. Then, processing the results in an appropriate way should help to confirm, refine and perhaps make more precise the correspondence between the two sets EU-DU-C and CN-C.

Potentially relevant terminology (significant words) from the field of Nuclear Non Proliferation (NP) was identified in two different ways, one for single words and another one for multi-word terms. For single words, a statistical method was used to identify which words are statistically significantly more frequent in NP documents compared to general documents. For this purpose a frequency list of words in a collection of NP documents was produced and compared to a generic word frequency list, using the standard 100 Million words British National Corpus BNC [11]. Both frequency lists were compared using the log-likelihood test. This test produces a ranked list of words that are surprisingly frequent. Figure 4 shows an example, the result obtained by ranking the words that describe items in section 0B001 of the Regulation.

To identify multi-word terms that are typical for NP-related texts, we first applied linguistic patterns to the NP texts to select noun phrases such as 'Word1 Word 2', 'Word4 Word4 Word5' or 'Word6 Word 7 Word8 Word9' and we then used statistical methods to select those that seemed most typical for texts from the NP domain. For this purpose, we used the "Tree Tagger" [12] software to recognise the part-of-speech of words (noun, verb, adjective, preposition, etc.) and then filtered out noun-noun or adjective-noun sequences, etc. As the resulting list contains very common noun phrases such as "current output" as well as specialist terminology such as "isotope separation", we applied various statistical measures (Mutual Information, log-likelihood test, etc.) to determine which combinations are statistically outstanding.

The overall results were thus NP-related lists of single words or compound expressions.

Having determined our words of significance, we illustrate how to use them to retrieve CN/TARIC items in correspondence to dual-use items of interest.

In this example we focus on item 0B001b11 whose textual description in the Council Regulation is:

*Centrifuge housing/recipients to contain the rotor tube assembly of a gas centrifuge, consisting of a rigid cylinder of wall thickness up to 30 mm with precision machined ends and made of or protected by materials resistant to corrosion by UF6”;*

where the words of significance appear underlined. They are displayed below according to the order of apparition in the text. In parenthesis, we have indicated their ranking by relative frequency, as explained above and as shown in Figure 4.

- centrifuge (3)
- gas centrifuge
- rotor tube assembly
- rotor (4)
- wall thickness
- cylinder (5)
- resistant (1)
- corrosion (2)

Goods: centrifuges	
Taric code	Description
<b>8421</b>	<b>Centrifuges, including centrifugal dryers; filtering or purifying machinery and apparatus, for liquids or gases</b>
1 <a href="#">8421000000</a>	Centrifuges, including centrifugal dryers; filtering or purifying machinery and apparatus, for liquids or gases
2 <a href="#">8421110000</a>	Centrifuges, including centrifugal dryers
3 <a href="#">8421192000</a>	Centrifuges of a kind used in laboratories
4 <a href="#">8421910000</a>	Of centrifuges, including centrifugal dryers

Figure 5: TARIC search hits on 'centrifuges'.

Goods: centrifugal	
Taric code	Description
<b>8413</b>	<b>Pumps for liquids, whether or not fitted with a measuring device; liquid elevators</b>
1 <a href="#">8413700000</a>	Other centrifugal pumps
2 <a href="#">8413708100</a>	Other centrifugal pumps
<b>8414</b>	<b>Air or vacuum pumps, air or other gas compressors and fans; ventilating or recycling hoods incorporating a fan, whether or not fitted with filters</b>
3 <a href="#">8414594000</a>	Centrifugal fans
<b>8421</b>	<b>Centrifuges, including centrifugal dryers; filtering or purifying machinery and apparatus, for liquids or gases</b>
4 <a href="#">8421000000</a>	Centrifuges, including centrifugal dryers; filtering or purifying machinery and apparatus, for liquids or gases
5 <a href="#">8421110000</a>	Centrifuges, including centrifugal dryers
6 <a href="#">8421910000</a>	Of centrifuges, including centrifugal dryers
<b>8450</b>	<b>Household or laundry-type washing machines, including machines which both wash and dry</b>
7 <a href="#">8450120000</a>	Other machines, with built-in centrifugal drier

Figure 6: TARIC search hits on 'centrifugal'.

CN/TARIC is then queried on each of these significant words, or part of them. For instance, the root word '**centrifug**' has four entries in TARIC: *centrifuge*, *centrifuges*, *centrifugal* and *centrifugation*. For the words *centrifuge* and *centrifugation*, no entry was found and no TARIC item identified. For the two other words, *centrifuges* and *centrifugal*, 4 and 7 entries were found respectively. It should also be mentioned that more than one significant word can be present in the description of a single CN item (e.g. 8421910000 in Figure 5 contains both *centrifuge* and *centrifugal*). 8 different results are therefore reported.

The result for the other significant words are:

- **rotor** with two entries (*rotor* and *rotors*) provide 4 and 2 hits respectively (total of 6).
- **cylinder** with three entries (*cylinder*, *cylinders*, *cylindrical*) provide 66, 13 and 27 hits respectively (104 in total).
- **corrosion** with *corrosion* and *corrosive* provides 5 and 1 hits.
- **wall thickness** gives 16 hits all together.

No result was found for *rotor(s) tube(s) assembly(ies)* nor with *gas centrifuge(s)*.

This short example search highlights a few specific points.

First of all, for one given item (in our case 0B001b11) the number of hits in CN-C can reach very high values (134 hits were obtained). Some of the CN-C items found can appear twice under two different entries. If we intend to search over a number of 50-100 different items as explained in section 4, the order of magnitude of hits to be processed will typically reach a few thousands.

A possible way to reduce the number of hits is to filter manually the significant words automatically extracted, either eliminating them from the list, either choosing the multiword expression(s) that include the term. For instance we can replace "resistant" by "materials resistant" or instead of searching for "cylinder" we can search for "rigid cylinder" and "tube cylinder". This will reduce the recall but could influence as well the precision of retrieval.

## 6. Discussion

**Classification schemes.** The classification used in the INFCIRC 254, as well as the one used in the Council Regulation on dual-use follow a logic very close to the description of the nuclear fuel cycle. Items go along a line of fuel cycle: enrichment (0B001), transfer of UF6 (0B002), conversion (0B003), Heavy Water production (0B004), fuel fabrication (0B005), reprocessing (0B006), Plutonium conversion (0B007), etc. In each of these chapters, a few sensitive items are selected and characterized.

From the customs point of view, the goods' classification system used in CN/TARIC is a type of 'partition' of the space of goods that are subject to trade. This means that, a given item (whether dual-use or not) is classified in only one way in a tree-like system. Typically, specialists in goods nomenclature establish a categorisation in a top-down fashion.

On the other hand, an exporter has to face the inverse problem: which CN-C value to assign to a given item in order to fit with the spirit of the classification system?

A further issue to keep in mind is that the 'words of significance' in the nuclear dual-use and in the CN system do not always match due to a different culture in the choice of terms.

**Making the best use of results.** For one given EU-DU item, the search over the significant words provides up to a few hundreds CN-C items. If we intend to search over 50-100 priority items as explained in section 4, the order of magnitude of hits to be verified will typically reach a few thousands. In order to get to the CN-C of interest, one has to eliminate all the irrelevant ones.

One way of helping this is to restrict the search to those CN chapters of interest. For example, some chapters can be excluded, such as 'Live animals; animal products' (chapter 1 to 5). In a first trial we have reduced the number of chapters to be searched by 39 chapters over a total of 97.

Another way to eliminate the irrelevant CN codes is to check one by one the CN items detected against the DU definition. This means that each given EU-DU item will be checked hundreds of time against each of the hit CN-C, until all the unnecessary CN-C are eliminated. The result is a list of possible CN-C for the given DU item.

A totally different approach would be to tackle the EU-DU classification problem by revising and embedding in the Combined Nomenclature ad-hoc CN-C codes, codes that refer to DU items in a non ambiguous way ideally leading to a clean one-to-one correspondence between EU-DU-C and CN-C.

**Use of languages: translations.** Another issue is the place and the role of translations. The translation of texts (EU-DU-C or CN-C) is not an exact science: a given word can be translated in more than one way depending on its context. Then, matching words can introduce some approximation or be incomplete, whilst the sense will continuously be matched.

For example, the word *diameter* (English) hits 145 items in the English version of TARIC, the word *diamètre* (French) hits 151 items in its French version and *diametro* hits 155 items in the Italian version.

This may decrease the probability of catching the 'right' CN-C items for a given DU item and it also makes the results dependent on the working language in which the search is performed.

**Synonyms.** Further, one should take into account the existence of synonyms. The words *tube* and *pipe*, *cylinder* or *tank* can refer to the same reality. This analysis can be assisted in an automatic way by the use of dictionaries of synonyms -preferably made by nuclear specialists and not general ones. Taking synonyms into account is expected to increase the number of hits (the 'recall') although one can hope to increase the retrieval precision as well.

In the end, we feel that the evaluation of the results by a domain expert will still be needed for deciding between relevant and irrelevant items.

**Steps ahead ?** Language technology has proven to be effective to filter and select 'free texts' that are originated in the open source. On the other hand, the CN definitions and their very structured organization cannot be ignored when trying to match DU descriptions. Working upstream, together with CN specialists can be a way to explore.

## 7. References

- [1] Nuclear Suppliers Group Guidelines (INFCIRC/254/Part 1)  
<http://www.nuclearsuppliersgroup.org/PDF/infirc254r8p1-060320.pdf>
- [2] Nuclear Suppliers Group Guidelines (INFCIRC/254/Part 2)  
<http://www.nuclearsuppliersgroup.org/PDF/infirc254r7p2-060320.pdf>
- [3] Council Regulation No 1334/2000 setting up a Community regime for the control of exports of dual-use items and technology, [http://trade.ec.europa.eu/doclib/docs/2004/february/tradoc\\_111418.pdf](http://trade.ec.europa.eu/doclib/docs/2004/february/tradoc_111418.pdf)
- [4] Council Regulation No 394/2006 amending and updating Regulation No 1334/2000  
[http://trade.ec.europa.eu/doclib/docs/2006/march/tradoc\\_127868.pdf](http://trade.ec.europa.eu/doclib/docs/2006/march/tradoc_127868.pdf)
- [5] Easy COMEXT site <http://fd.comext.eurostat.cec.eu.int/xtweb/mainxtnet.do>
- [6] COMTRADE site <http://unstats.un.org/unsd/comtrade/>
- [7] Access point to MetaData resources, included the Combined Nomenclature  
[http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP\\_PUB\\_WELC](http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC)
- [8] TARIC Regulation,  
[http://ec.europa.eu/taxation\\_customs/dds/premnotes/2003\\_04\\_30\\_JOC103\\_EN.pdf](http://ec.europa.eu/taxation_customs/dds/premnotes/2003_04_30_JOC103_EN.pdf)
- [9] TARIC consultation site, [http://ec.europa.eu/taxation\\_customs/dds/en/tarhome.htm](http://ec.europa.eu/taxation_customs/dds/en/tarhome.htm)
- [10] Ignat Camelia, Bruno Pouliquen, Ralf Steinberger & Tomaž Erjavec (2005). A tool set for the quick and efficient exploration of large document collections. Proceedings of the Symposium on Safeguards and Nuclear Material Management. 27th Annual Meeting of the European Safeguards Research and Development Association (ESARDA-2005). London, UK, 10-12 May 2005
- [11] The British National Corpus home page: <http://www.natcorp.ox.ac.uk/corpus/index.xml>
- [12] TreeTagger - a language independent part-of-speech tagger: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>