An XML-UNL Model for Knowledge-Based Annotation

Jesús Cardeñosa, Carolina Gallardo and Luis Iraola

Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Madrid, Spain
{carde,carolina,luis}@opera.dia.fi.upm.es

Abstract. Efficient document search and description has radically changed with the widespread availability of electronic documents through Internet. Nowadays, efficient information search systems require to go beyond HTML-annotated documents. Complex information extraction tasks require to enrich text with semantic annotations that allow deeper and more detailed content analysis. For that purpose, new labels or annotations need to be defined. In this paper we propose to use UNL, an interlingua defined by the United Nations University, as a language neutral standard content representation in Internet. The use of UNL would open documents to a new dimension of semantic analysis, thus overcoming the limitations of current text-based analysis techniques.

1 Introduction

XML [1] is an standardized annotation language currently employed for a variety of purposes. For any given domain, the set of tags defined in its DTD attempts to capture the logical content structure of typical documents of the domain. So annotated, documents can be exploited by sophisticated document management systems that provide precise answers to users' queries. One the most promising uses of XML is the possibility of replacing textual document bases by their XML counterparts for document management purposes as well as for content management.

The capability of the XML standard to define the different information items present in a given document facilitates subsequent information extraction operations. This capability makes XML an ideal choice for annotating text corpora.

Annotated corpora have been one of the most useful resources in the last years for the study of linguistic phenomena. This orientation towards linguistic analysis has frequently associated corpus annotation with tasks such as part of speech tagging, chunking and parsing.. The Brown Corpus [2] or the British National Corpus [3] are examples of such annotated corpora. This sort of annotation is useful for many purposes but may be insufficient for information management tasks and for the location of very specific information items.

Corpus annotation poses significant difficulties when the goal is the representation and classification of information expressed in text form. While one could say that lexical and syntactic annotation of textual corpora is a more or less solved problem, semantic tagging is still a challenging goal currently aimed by several research lines.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 300–308.

Semantic corpora annotation has traditionally focused on the tasks of sense disambiguation of linguistic expressions [4] [5], definition of the conceptual relations between a heading verb and the dependent elements within the sentence [6], and frequently on tagging and classification of key concepts and specific elements pertaining to a given domain, like in [7] and [8]. Therefore, content analysis and representation by semantic tagging has in most cases a descriptive character and semantic annotation is mostly driven by the specific terminology of the domain. In multilingual corpora, semantic annotation is more superficial, and practically stops at the linguistic level.

A departure from this approach (domain and language dependency for semantic tagging) is one where textual information is expressed in a language-independent formalism whose semantic relations do not depend on any specific given domain. Such language independent formalisms are known as interlinguas in the field of Machine Translation.

An interlingua is an artificial language able to represent meaning in a language-independent way. Since one of the purposes of XML tagging is semantic annotation of the informational contents of a given document, there is in principle no special objection in applying XML tagging to represent a document written in an interlingua. The interlingua approach is not new and its origins can be traced back to the late eighties, when a number of multilingual machine translation systems were designed and implemented, such as Pivot [9] and Atlas-II [10]. In the nineties, machine translation systems evolved into the so-called knowledge-based machine translation systems, of which Kant [11] and Mikrokosmos [12] are two prominent examples.

The scalability problems of interlingua-based multilingual translation systems almost led to the rejection of the concept of interlingua. However, in 1996 the Institute of Advanced Studies of the United Nations University launched a new research project that rescued the interlingua approach for supporting multilingual content exchange in Internet by means of the use of UNL (Universal Networking Language).

UNL can be viewed as a reincarnation of the traditional concept of interlingua as an intermediate abstract representation common to all natural languages in a multilingual machine translation system. But UNL goes beyond the notion of a classical interlingua: it also serves for representing informational contents in any domain and in a language independent manner. UNL is endowed with an expressive capability similar to a natural language but with the features of a formalized language; its syntax and semantics are well defined, so UNL may be employed in information extraction and reasoning tasks.

2 The UNL System

UNL is an artificial language designed to represent textual content written in any natural language. The specifications of the UNL [13] formally define the language and its components. These are basically the following ones:

Universal words. They conform the vocabulary of the language, i.e., they can be considered the lexical items of UNL. In order to be able to express any concept occurring in a natural language, UNL proposes the usage of English headwords possibly modified by a series of semantic restrictions that eliminate potential ambiguities of those headwords. When there is no English headword suitable to express the intended concept, UNL allows the usage of words coming from other languages. In this way, the interlingua achieves the same lexical richness than natural languages but without their ambiguity. Take, for example, the English word "construction" meaning "the action of constructing" and also the "final product or result of constructing". The basic universal word "construction" will be paired with two different restricted universal words:

$\texttt{construction}_1$: construction(icl>action)
$\texttt{construction}_2$: construction(icl>concrete thing)

where "icl" is the abbreviation for "included".

Relations. There are a set of 41 basic relations that allow for the definition of any possible semantic relation among concepts. They include argumentative (agent, object, goal), circumstantial (purpose, time, place) and logic relations (conjunction and disjunction). For example, in the sentence "The boy eats potatoes in the kitchen", there is a main predicate ("eats") and three relations, two of them are instances of argumentative relations ("boy" is the agent of the predicate "eats", whereas "potatoes" is its object) and one circumstantial relation ("kitchen", the physical place where the action described in the sentence takes place).

Attributes. They express several types of semantic information that modify the relations and/or the universal words employed for expressing the content of a given text. This information includes time and aspect of the event, negation and modality of predication, type of reference of the entities described by the universal words, number and/or gender, etc. In the previous sentence, attributes are needed to express plurality in the object ("potatoes"), definite reference in the both the agent ("boy") and the place ("kitchen") and finally and special attribute denoting which UW is the head of the whole expression (the entry node).

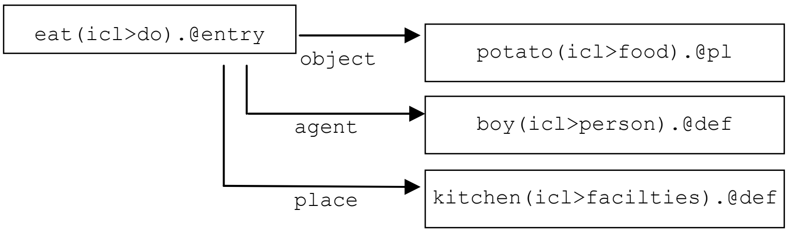


Fig 1. Graphical representation of a UNL expression.

Formally, a UNL expression has the form of a semantic net, where the nodes (universal words) are linked by arcs labeled with the UNL conceptual relations. The graphical representation of the sentence "the boy eats potatoes in the kitchen" in UNL is shown in figure 1, whereas its representation in the UNL syntax is as follows:

```
agt(eat(icl>do).@entry, boy(icl>person).@def)
obj(eat(icl>do).@entry, potato(icl>food).@pl )
plc(eat(icl>do).@entry, kitchen(icl>facilities).@def)
```

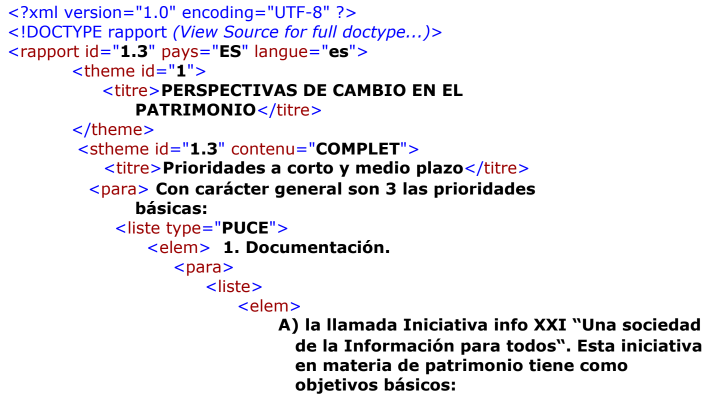
The capabilities of UNL for representing content independently from the source language led the authors to participate in the Herein system and to use UNL for supporting multilingual services in this particular system.

3 The UNL Approach in Herein

The Herein system (IST-2000-29355) [14] is a perfect example of a massively multilingual environment. It constitutes an Internet-based facility for improving cultural heritage management methods at the European level. Among the main tasks of the project, participant countries must compose a report providing detailed information about all aspects regarding cultural heritage.

Due to the large number of countries participating in the project (almost thirty) and the huge variety of topics that comprises cultural heritage (legislation, preservation, dissemination, etc.), there was an urgent need to standardize both the format and the structure of the contents that each country should provide. A definite structure was established and every country involved in Herein had to integrate its particular contents into such structure. Eventually, this structure turned out to be a de-facto standard for the description of the cultural heritage issues of a country.

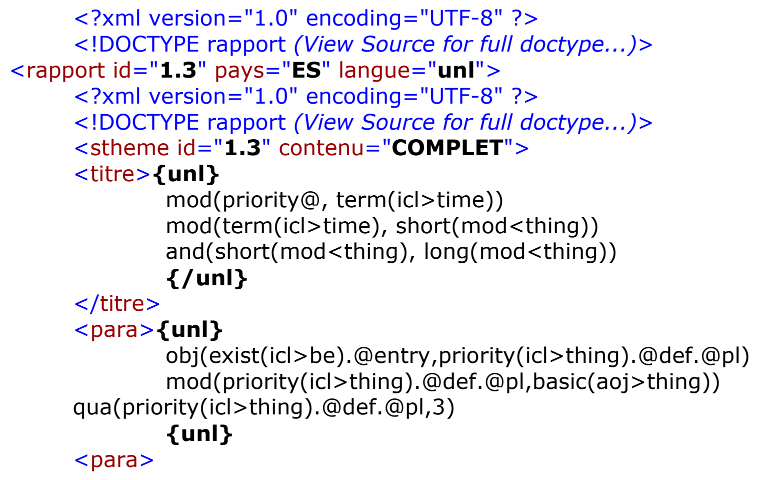
The supporting format chosen for the structured reports on cultural heritage of each participating country was XML. Figure 2 shows the appearance of a typical report in the Herein project: a fragment extracted from the Spanish Report.



```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE rapport (View Source for full doctype...)>
<rapport id="1.3" pays="ES" langue="es">
        <theme id="1">
            <titre>PERSPECTIVAS DE CAMBIO EN EL
                PATRIMONIO</titre>
        </theme>
         <stheme id="1.3" contenu="COMPLET">
            <titre>Prioridades a corto y medio plazo</titre>
          <para> Con carácter general son 3 las prioridades
                básicas:
             <liste type="PUCE">
                 <elem>  1. Documentación.
                     <para>
                         <liste>
                             <elem>
                                  A) la llamada Iniciativa info XXI "Una sociedad
                                    de la Información para todos". Esta iniciativa
                                    en materia de patrimonio tiene como
                                    objetivos básicos:
```

Fig 2. Example of Spanish content in XML structure

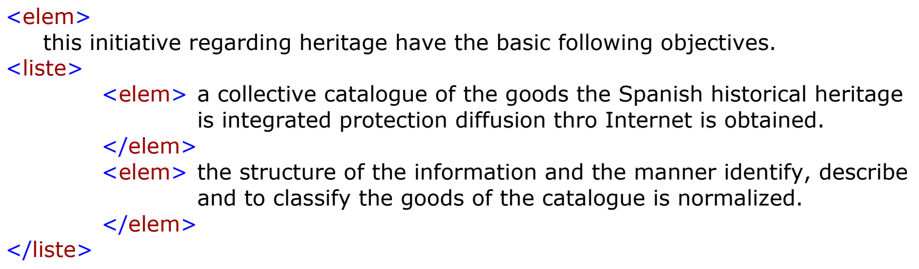
The complete report of the Spanish cultural contents was codified into UNL as an initiative of the Spanish government, representative institution of the Herein contents in the Spanish language, and in collaboration with the Spanish Language Center, representative and responsible of the Spanish language in the UNL program.



```
          <?xml version="1.0" encoding="UTF-8" ?>
          <!DOCTYPE rapport (View Source for full doctype...)>
      <rapport id="1.3" pays="ES" langue="unl">
          <?xml version="1.0" encoding="UTF-8" ?>
          <!DOCTYPE rapport (View Source for full doctype...)>
          <stheme id="1.3" contenu="COMPLET">
          <titre>{unl}
                  mod(priority@, term(icl>time))
                  mod(term(icl>time), short(mod<thing))
                  and(short(mod<thing), long(mod<thing))
                  {/unl}
          </titre>
          <para>{unl}
                  obj(exist(icl>be).@entry,priority(icl>thing).@def.@pl)
                  mod(priority(icl>thing).@def.@pl,basic(aoj>thing))
          qua(priority(icl>thing).@def.@pl,3)
                  {unl}
          <para>
```

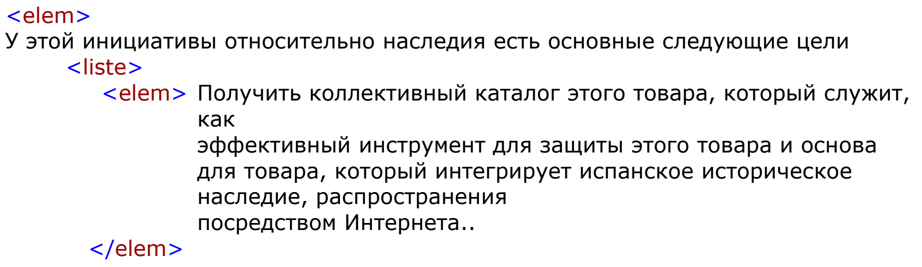
Fig 3. UNL code embedded into an XML document.

The UNL code has been embedded into the XML structure shared by all reports, as if the UNL code were another natural language (see figure 3). The difference lies in the fact that the aforementioned code can be extracted from the XML file and employed by the natural language generator of any language. After generation [15], the corresponding text will be inserted into the XML structure of the document. The result is shown in figures 4 and 5 for the English and Russian language generators.



```
<elem>
   this initiative regarding heritage have the basic following objectives.
<liste>
        <elem> a collective catalogue of the goods the Spanish historical heritage
               is integrated protection diffusion thro Internet is obtained.
        </elem>
        <elem> the structure of the information and the manner identify, describe
               and to classify the goods of the catalogue is normalized.
        </elem>
</liste>
```

Fig 4. Output text of the English generator



```
<elem>
У этой инициативы относительно наследия есть основные следующие цели
     <liste>
        <elem> Получить коллективный каталог этого товара, который служит,
               как
               эффективный инструмент для защиты этого товара и основа
               для товара, который интегрирует испанское историческое
               наследие, распространения
               посредством Интернета..
       </elem>
```

Fig 5. Output text of the Russian generator

The complete integration of UNL into the Herein system is illustrated in figure 6. In this figure, it can be seen how an original XML document about Spanish heritage is the input to an UNL editor once its XML tags have been removed from it and the textual content extracted. The UNL editor is a tool that enables its user to encode Spanish sentences into UNL expressions. The degree of automation depends on the current state of Spanish-UNL dictionaries and its syntactic and semantic analyzers. The output of the UNL editor is a plain document written in UNL (that is, no XML tagging is present). This UNL document goes directly into the language generators, for example the English and Russian language generators. These generators yield the contents of the original XML Spanish document but now in English and Russian respectively. The final step is the "XMLization" of these plain documents according the DTD adopted in the Herein system.

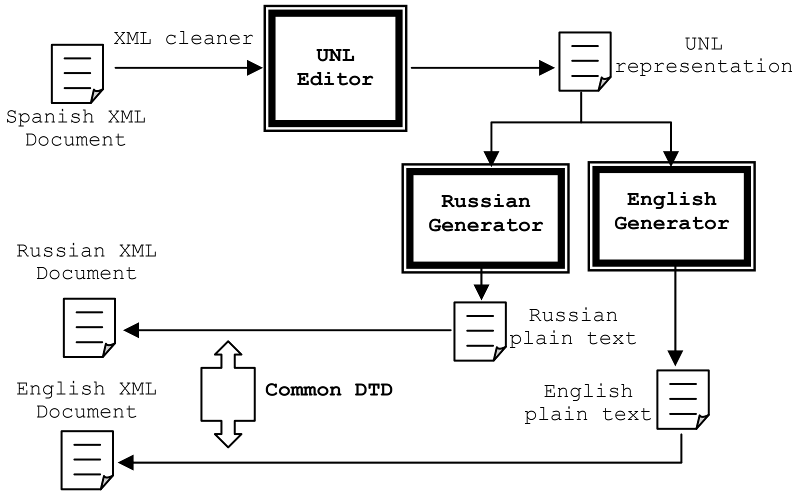


Fig 6. Model for the Integration of UNL into Herein

Within the Herein system, UNL has been integrated with XML mainly for the support and maintenance of multilingual documents. However, the integration of UNL into XML can be further explored in order to take further advantage of the UNL code for semantic annotation.

4 Knowledge-Based Annotation in XML: a Three-Dimensional Approach

Currently, a closer integration of UNL and XML is being studied from a different perspective [16] but within the same framework here described. This innovative work attempts to define environments and architectures that allow the inclusion of XML tags that identify individual UNL elements (i.e. universal words, relations and attributes). This fine-grained semantic representation will pave the way to more intelligent information extraction tasks. This is possibly the most immediate research line that would produce an effective integration of XML and UNL. If we are able to define a suitable XML syntax for representing UNL, and also to semantically annotate the content of a document not only according to an set of domain-specific descriptive terms but also using the semantic relations that connect the concepts present in the document, we will transform a "one-dimensional" textual document into a "three-dimensional" document.

Why a third dimension? We may consider the text as the first dimension of a document. It is the basis of any linguistic analysis and it is certainly the basis of the encoding process in the UNL system. Layout, formatting and hyper-linking constitute a second dimension of the document. This second dimension provides cues about the specific information pieces contained in a document and facilitates searching and extraction. But, if in addition to the first and second dimensions we are able to capture the semantic relations among the concepts present in the document, we may say that a third dimension has been made available, a dimension where the knowledge contained in a given document is made explicit. Document management systems become knowledge management systems by exploiting this third dimension, implementing knowledge-based reasoning procedures able to produce intelligent answers to complex queries.

The integration of the UNL representation will improve the quality and depth of the knowledge expressed by XML tagging. The UNL relations are based on what has been traditionally known as conceptual or thematic relations or simply cases. Along this line, other authors are using these relations as the leitmotiv for semantic annotation [6]. However, at this point some reflections should be made about the nature of UNL, as they back UNL as a firm candidate for the task of representing the knowledge level in any XML document. Key UNL characteristics are:

(a) The set of necessary relations existing between concepts is already standardized [13]. This is the result of intensive research on the thematic roles existing in natural languages by a number of experts in the area of MT and AI.

(b) Similarly, the set of necessary attributes that modify concepts and relations is fixed and well-defined.

(c) The UNL syntax and semantics are formally defined, UNL can be viewed as a formalism for representing knowledge.

In short, UNL has in its favor the standardization of the process of representing knowledge coming from documents written in a natural language. In the following example, we show the approach to be followed along this direction. We will show an example of the abovementioned third dimension applied to a paragraph extracted from the Herein Spanish report (originally in Spanish but here in English for readability reasons):

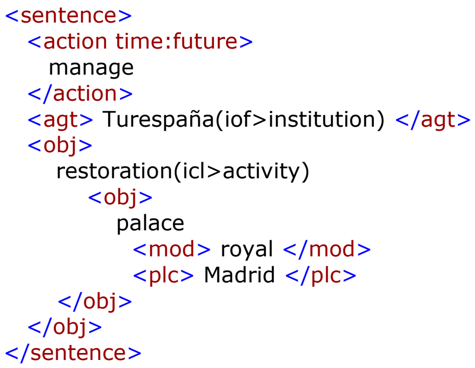
```
<para> The restoration of the Royal Palace of Madrid
will be managed by Turespaña. </para>
```

Its UNL representation is as follows:

```
agt(manage(icl>do).@entry.@future,
    Turespaña(iof>institution))
obj(manage(icl>do).@entry.@future,
     restoration(icl>activity).@def)
obj(restoration(icl>activity).@def,
     palace(icl>building).@def)
mod(palace(icl>building).@def, royal(mod<thing))
plc(palace(icl>building).@def, Madrid(iof>city))
```

The encoded meaning is that of an action carried out by an agent (agt) and described as a managing action performed by the institution named (iof) 'Turespaña'. The object (obj) of the managing action is a restoration activity. It is also specified that the object of the restoration is a palace, a type of building (icl), modified (mod) by the property of being a royal palace and located (plc) in Madrid. Additionally, the time of the action is future.

It is clearly possible to define an XML-based tag language for expressing the elements of a UNL representation: UNL relations could be considered as XML tags, attributes could be represented as XML attributes and universal words may just be textual data enclosed within UNL relation tags. Figure 7 presents the previous UNL representation along these lines.



```
<sentence>
  <action time:future>
    manage
  </action>
  <agt> Turespaña(iof>institution) </agt>
  <obj>
     restoration(icl>activity)
        <obj>
           palace
             <mod> royal </mod>
             <plc> Madrid </plc>
     </obj>
  </obj>
</sentence>
```

Fig 7. UNL representation using XML-based tags.

This representation conforms to a very precise characterization of the semantic relations and the concepts present in the sentence. Therefore, the knowledge implicit in the sentence has been explicitly formalized and integrated within an XML-based document structure.

5 Conclusions

In this paper we have presented a new approach for representing knowledge contained in textual documents using an interlingua. The use of UNL allows and facilitates the integration of knowledge into an XML structure by means of the definition of a set of XML-based tags and attributes suited for the basic elements of a UNL representation. At the moment we are testing the adequacy of UNL representations embedded into XML documents for information extraction tasks. We are also devising an interactive system of queries over contents so represented. Our approach may prove useful for annotating multilingual text corpora with semantic information, thus extending the range of applications of an interlingua originally designed for multilingual generation purposes.

References

1. W3C Recommendation. *Extensible Markup Language*, www.w3.org/TR/2004/REC-xml-20040204/, 2004.
2. Francis W.N. and Kucer H. *Brown Corpus Manual*, helmer.aksis.uib.no/icame/brown/bcm.html, 1979.
3. Guy A., and Burnard L *The BNC Handbook - Exploring the British National Corpus with SARA*. Edinburgh, UK. Edinburgh University Press, 1998.
4. Garside, R. and Rayson, P. Higher-level annotation tools, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, 1997.
5. Thomas, J., and Wilson, A. Methodologies for studying a corpus of doctor-patient interaction. In J. Thomas and M. Short (eds.). *Using corpora for language research*. Longman, London, 1996.
6. Kingsbury M., Palmer M., and Marcus M. In Proceedings of the Human Language Technology Conference, San Diego, California, 2002.
7. Ohta T., Tateisi Y., Takai Y. and Tsujii J. A Semantically Annotated Corpus from MEDLINE Abstracts. In the Proceedings of Genome Informatics. Tokyo, Japan. Universal Academy Press Inc., 1999.
8. Vintar S. and Volk M. Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval. In Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM), Cavtat-Dubrovnik, Croatia, 2003.
9. Muraki, K.: PIVOT. Two-phase machine translation system. Proceedings of the Second Machine Translation Summit. Tokyo, Japan, 1989.
10. Uchida, H.: ATLAS-II. A machine translation system using conceptual structure as an interlingua. Proceedings of the Second Machine Translation Summit. Tokyo, Japan, 1989.
11. Nyberg E. and Mitamura T. The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains. Proceedings of COLING-92: 15th International Conference on Computational Linguistics. Nantes, France, 1992.
12. Beale, S., S. Nirenburg and K. Mahesh. Semantic Analysis in the Mikrokosmos Machine Translation Project. Proceedings of the 2nd Symposium on Natural Language Processing. Bangkok, Thailand, 1995.
13. Uchida, H.: The Universal Networking Language. Specifications. www.undl.org, 2002.
14. HEREIN Project (IST-2000-29355): Final Report. European Commission, 2003.
15. Cardeñosa J., Gallardo C and Tovar E. Standardization of the generation process in a multilingual environment. In Proceedings of the International Conference Convergences 03. December 2003. Alexandria. Egypt, 2003.
16. Wang-Ju. Work in progress accessible from: www-clips.imag.fr/geta/wang-ju.tsai/viewer/Multiple_js.htm, 2004.