The Wikipedia XML Corpus

Ludovic Denoyer, Patrick Gallinari
14th of April, 2006

Laboratoire d'Informatique de Paris 6
8 rue du capitaine Scott
75015 Paris

http://www-connex.lip6.fr/denoyer/wikipediaXML
{ludovic.denoyer, patrick.gallinari}@lip6.fr

1 Introduction

Wikipedia[1] is a well know free content, multilingual encyclopedia written collaboratively by contributors around the world. Anybody can edit an article using a wiki markup language that offers a simplified alternative to HTML. This encyclopedia is composed of millions of articles in different languages.

Content-oriented XML retrieval is an area of Information Retrieval (IR) research that is receiving an increasing interest. There already exists a very active community in the IR/ XML domain which started to work on XML search engines and XML textual data. This community is mainly organized since 2002 around the INEX initiative (INitiative for the Evaluation of XML Retrieval) which is funded by the DELOS network of excellence on Digital Libraries.

In this article, we describe a set of XML collections based on Wikipedia. These collections can be used in a large variety of XML IR/Machine Learning tasks like ad-hoc retrieval, categorization, clustering or structure mapping. These corpora are currently used for both, INEX 2006[2] and the XML Document Mining Challenge[3]. The article provides a description of the corpus.

The collections are downloadable on the website:

- *http://www-connex.lip6.fr/∼denoyer/wikipediaXML*

2 Description of the corpus

The corpus is composed of 8 main collections corresponding to 8 different languages[4] : English, French, German, Dutch, Spanish, Chinese, Arabian and Japanese. Each collection is a set of XML documents built using Wikipedia and encoded in UTF-8. In addition to these 8 collections, we also provide different *additional collections* for other IR/Machine Learning tasks like categorization and clustering, NLP, machine translation, multimedia IR, entity search, etc.

[1] http://www.wikipedia.org
[2] http://inex.is.informatik.uni-duisburg.de/2006
[3] http://xmlmining.lip6.fr
[4] Some additional languages will be added during the next months.

2.1 Main Collections

The main collections are a set of XML files in 8 different languages. The table 1 gives a detailed description of each collection.

Collection name	Language	Number of documents	Size of the collection (MegaBytes)
main-english	English	659,388	≈ 4,600
20060130_french	French	110,838	≈ 730
20060123_german	German	305,099	≈ 2,079
20060227_dutch	Dutch	125,004	≈ 607
20060130_spanish	Spanish	79,236	≈ 504
20060303_chinese	Chinese	56,661	≈ 360
20060326_arabian	Arabian	11,637	≈ 53
20060303_japanese	Japanese	187,492	≈ 1,425

Table 1. General statistics about the *Main Collections*

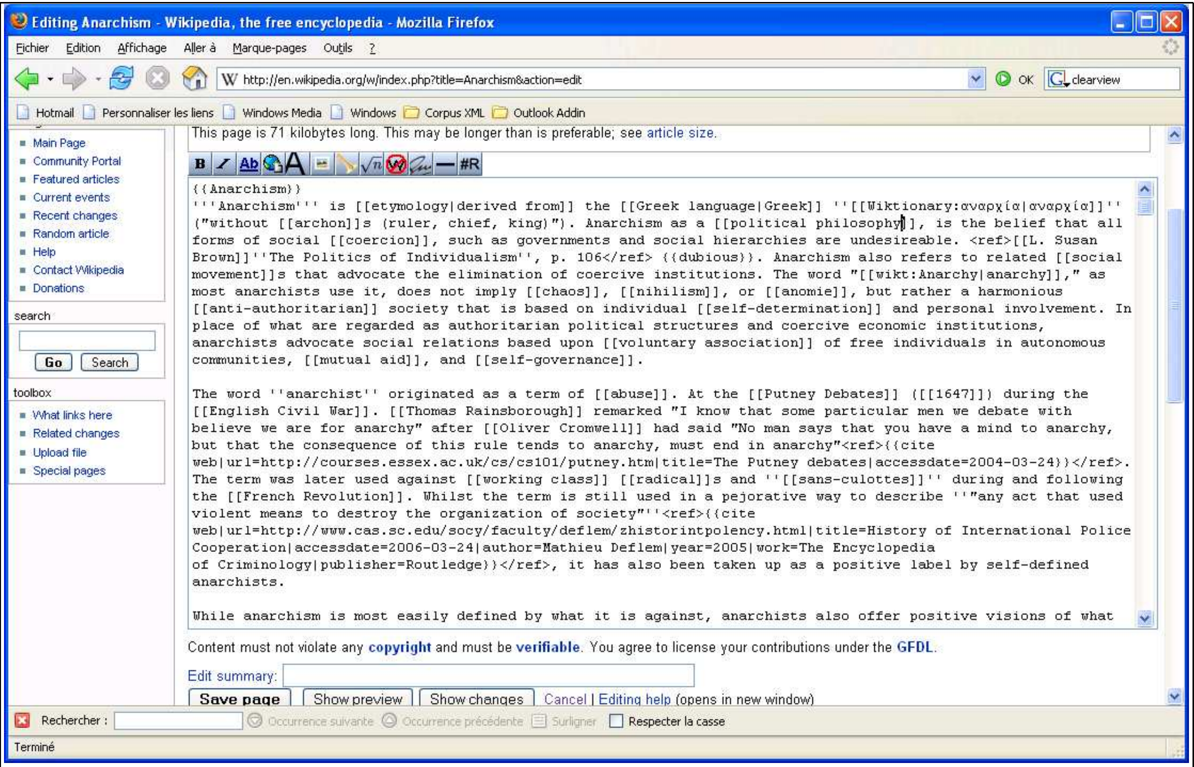
Each collection contains a set of documents where each filename is a number corresponding to the id of the file (for example : *15243.xml*). Each id is unique and each file corresponds to an article of Wikipedia. We only kept articles and removed all the wikipedia pages corresponding to ”‘Talks”’, ”‘Template”’, etc.. Each file is an UTF-8 document which is created from the wikitext of the original article. Figure 1 gives an example of an English article extracted from the corpus.

Tag labels We introduced different tags in order to represent the different parts of a document. We distinguish two types of tags:

- The general tags (*article,section, paragraph,....*) that do not depend on the language of the collection. These tags correspond to the structural information contained in the wikitext format (for example : == *Main part* == is transformed into <*title*>*Main part*< /*title*>)
- The template tags (*template_infobox,...*) represent the information contained into the wikipedia templates. Wikipedia templates are used to represent a repetitive type of information. For example, each country described into wikipedia starts with a table containing its population, language, size,... In order to uniformize this type of information, wikipedia uses templates. These templates are translated into XML using tags starting by *template_...* (for example : *template_country*). The template tags depend on the language of the collection because the templates are not the same depending on the language of the wikipedia collection used.

The DTD is downloadable on the Web site.

Statistics about the collections These statistics are given in table 2



This page is 71 kilobytes long. This may be longer than is preferable; see article size.

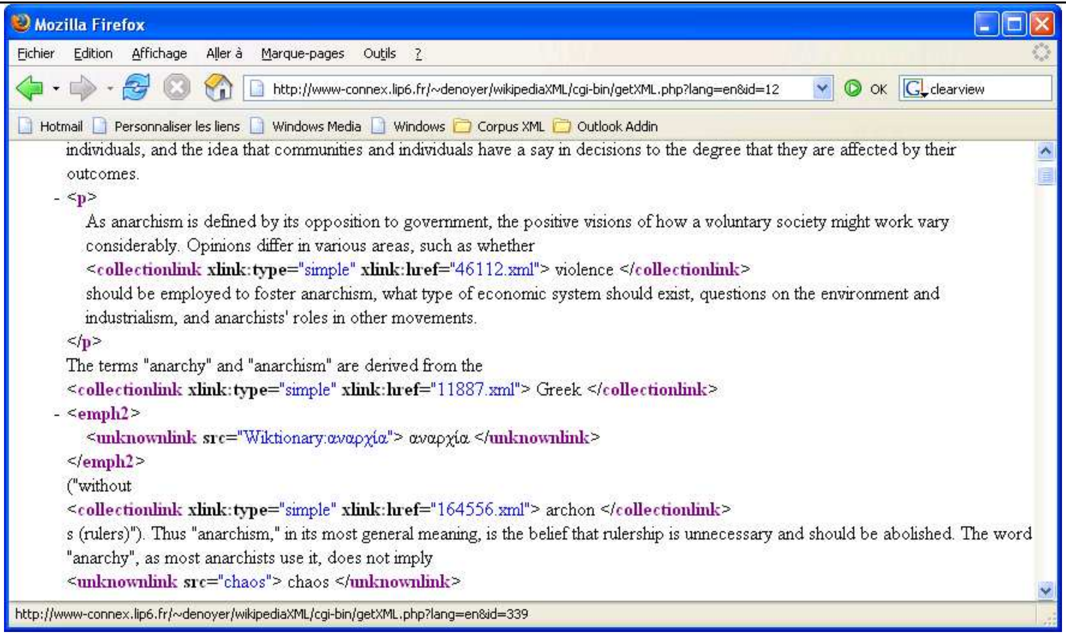
```
{{Anarchism}}
'''Anarchism''' is [[etymology|derived from]] the [[Greek language|Greek]] ''[[Wiktionary:αναρχία|αναρχία]]''
("without [[archon]]s (ruler, chief, king)"). Anarchism as a [[political philosophy]], is the belief that all
forms of social [[coercion]], such as governments and social hierarchies are undesireable. <ref>[[L. Susan
Brown]]''The Politics of Individualism'', p. 106</ref> {{dubious}}. Anarchism also refers to related [[social
movement]]s that advocate the elimination of coercive institutions. The word "[[wikt:Anarchy|anarchy]]," as
most anarchists use it, does not imply [[chaos]], [[nihilism]], or [[anomie]], but rather a harmonious
[[anti-authoritarian]] society that is based on individual [[self-determination]] and personal involvement. In
place of what are regarded as authoritarian political structures and coercive economic institutions,
anarchists advocate social relations based upon [[voluntary association]] of free individuals in autonomous
communities, [[mutual aid]], and [[self-governance]].

The word ''anarchist'' originated as a term of [[abuse]]. At the [[Putney Debates]] ([[1647]]) during the
[[English Civil War]]. [[Thomas Rainsborough]] remarked "I know that some particular men we debate with
believe we are for anarchy" after [[Oliver Cromwell]] had said "No man says that you have a mind to anarchy,
but that the consequence of this rule tends to anarchy, must end in anarchy"<ref>{{cite
web|url=http://courses.essex.ac.uk/cs/cs101/putney.htm|title=The Putney debates|accessdate=2004-03-24}}</ref>.
The term was later used against [[working class]] [[radical]]s and ''[[sans-culottes]]'' during and following
the [[French Revolution]]. Whilst the term is still used in a pejorative way to describe ''"any act that used
violent means to destroy the organization of society"''<ref>{{cite
web|url=http://www.cas.sc.edu/socy/faculty/deflem/zhistorintpolency.html|title=History of International Police
Cooperation|accessdate=2006-03-24|author=Mathieu Deflem|year=2005|work=The Encyclopedia
of Criminology|publisher=Routledge}}</ref>, it has also been taken up as a positive label by self-defined
anarchists.

While anarchism is most easily defined by what it is against, anarchists also offer positive visions of what
```

Content must not violate any copyright and must be verifiable. You agree to license your contributions under the GFDL.

the wiki text


Mozilla Firefox
Fichier Edition Affichage Aller à Marque-pages Outils ?
http://www-connex.lip6.fr/~denoyer/wikipediaXML/cgi-bin/getXML.php?lang=en&id=12
OK
clearview
Hotmail
Personnaliser les liens
Windows Media
Windows
Corpus XML
Outlook Addin
http://www-connex.lip6.fr/~denoyer/wikipediaXML/cgi-bin/getXML.php?lang=en&id=339

```
    individuals, and the idea that communities and individuals have a say in decisions to the degree that they are affected by their
    outcomes.
  - <p>
      As anarchism is defined by its opposition to government, the positive visions of how a voluntary society might work vary
      considerably. Opinions differ in various areas, such as whether
      <collectionlink xlink:type="simple" xlink:href="46112.xml"> violence </collectionlink>
      should be employed to foster anarchism, what type of economic system should exist, questions on the environment and
      industrialism, and anarchists' roles in other movements.
    </p>
    The terms "anarchy" and "anarchism" are derived from the
    <collectionlink xlink:type="simple" xlink:href="11887.xml"> Greek </collectionlink>
  - <emph2>
      <unknownlink src="Wiktionary:αναρχία"> αναρχία </unknownlink>
    </emph2>
    ("without
    <collectionlink xlink:type="simple" xlink:href="164556.xml"> archon </collectionlink>
    s (rulers)"). Thus "anarchism," in its most general meaning, is the belief that rulership is unnecessary and should be abolished. The word
    "anarchy", as most anarchists use it, does not imply
    <unknownlink src="chaos"> chaos </unknownlink>
```

The XML obtained

Fig. 1. Example of wiki → XML transformation for the *Anarchy* article (*12.xml*)

Language	Mean size of document (bytes)	Mean Document Depth	Number of Nodes/Document
English	7,261	6.72	161.35
French	6,902	7.07	175.54
German	7,106	6.76	151.99
Dutch	5,092	6.41	122.8
Spanish	6,669	6.65	165.74
Chinese	6,664	6.91	179.23
Arabian	4,826	5.85	182.1
Japanese	7,973	7.1	94.96

Table 2. Statistics about the structure of the documents from the *Main Collections*

2.2 Categories

The documents of the wikipedia XML collections are organized in a hierarchy of categories defined by the authors of the articles. For each main collection, we propose a set of files describing:

- the hierarchy of categories (file : categories_hcategories.csv)
- the categories of each articles (file : categories_categories.csv)
- the categories names (file : categories_name.csv)

Table 3 gives statistics about the categories.

Language	Number of categories in the hierarchy	Mean number of categories for each document
English	113,483	2.2849
French	28,600	1.9570
German	27,981	2.5840
Dutch	13,847	1.6628
Spanish	12,462	1.6180
Chinese	27,147	2.0797
Japanese	26,730	2.0039

Table 3. Statistics about the categories of the *Main Collections*

3 Additional collections

We also propose additional collections. These collections can be used in a large variety of Information Retrieval and Machine Learning tasks. Some other collections will be added in the future and are not described here.

3.1 Categorization/Clustering Collections

English Multi-Label Categorization Collection For the English collection, we also provide a *Multi-Label Categorization Collection* with :

- A list of articles from the *Main English Collection*
- A set of categories (without hierarchy, based on the *Portals* of wikipedia): a document belongs to one or more categories

This corpus can be used to compare categorization algorithms and it is described in table 4.

Number of Documents	415,310
Number of categories	72
Mean number of categories for each document	2.2
Mean number of documents for each category	12,5
Number of documents in the larger category	137,9
Number of Documents in the smaller category	108

Table 4. Statistics about the *English Multi-Label Categorization Collection*

English Single-Label Categorization Collection - XML Document Mining Collection We provide a specific collection where each document belongs to **exactly** one category. It is composed of the documents of the preceding collection belonging to a single category. This collection can be used for categorization and clustering of documents (see table 5). This collection is aimed at categorization/clustering benchmark.

Number of categories	60
Number of documents	150,094
Number of train documents	75,047
Number of test documents	75,047
Mean number of categories for each document	1
Structure of the corpus	The directory *documents* contains all the corresponding articles. The directory *relfiles* contains one file per category giving the id of the documents that belongs to this category[5].

Table 5. Statistics about the *XML Document Mining Challenge Collection* (*Single-Label Categorization Collection*)

3.2 Multimedia English Collection

This collection corresponds to the Main English Collection with the pictures of the different articles. This collection can be used for Multimedia Information Retrieval. Table 6 gives statistics about this collection.

Number of documents	659,388
Number of pictures	more than 300,000
Approximate size of the corpus	$\approx$ 60Gb

Table 6. Statistics about the *Multimedia English Corpus*

3.3 Entity corpus

We provide an *Entity Corpus* where each article of the *Main English Corpus* has been tagged using a set of possible entity types extracted using the different categories of wikipedia. For example : *Silverster Stallone* has been tagged as <*actors*>*Silverster Stallone*< /*actors*>. Table 7 gives statistics about this collection.

Number of Documents	XML Entity collection
Number of documents	659,388
Size of the corpus	$\approx$ 6 Gb

Table 7. Entity Collections

4 Conclusion

This technical report describes XML collections based on Wikipedia and developed for Structured Information Retrieval, Structured Machine Learning and Natural Language processing. Other collections will be added in the future.

5 Acknowlegment

The wikipediaXML corpus is distributed under the GPL Documentation license. It is completely free and can be used for non-profit educational and research purposes. All publications based on the wikipediaXML corpus should cite this technical report.