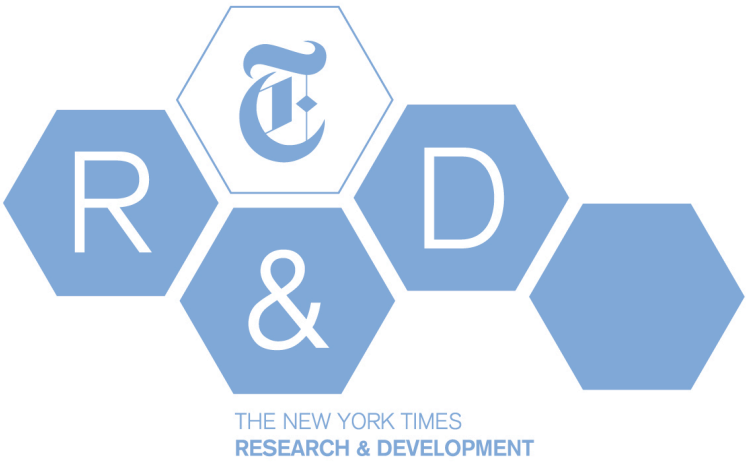

New York Times Corpus
Corpus Overview

Prepared By: Evan Sandhaus

New York Times, Research and Development
620 8th Ave 28th Floor
New York, NY 10018

1. Introduction

The purpose of this document is to provide an overview of the New York Times Corpus. The corpus is drawn from the historical archive of the New York Times and includes metadata provided by the New York Times Newsroom, the New York Times Indexing Service and the online production staff at nytimes.com. This corpus contains nearly every article published in the New York Times between January 01, 1987 and June 19th 2007. However, articles from wire services that appeared in the New York Times during this period are not included.

This document starts with an explanation of the contents and structure of the corpus' documents. Following that, this document presents an overview of the New York Time production process to provide context for understanding the contents of corpus. This document concludes with a number of useful statistics about the corpus.

2. Document Content and Structure

The New York Times corpus is provided as a collection of XML documents that conform to version 3.3 of the News Industry Text Format (NITF) specification. For more information on the NITF specification please visit http://www.nitf.org. Figure 1 shows a sample New York Times Corpus Document. Table 1 provides a brief explanation of each data field in the sample document. Sections 2.2.1 through 2.2.48 provide detailed descriptions of each data field.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nitf SYSTEM "http://www.nitf.org/IPTC/NITF/3.3/specification/dtd/nitf-3-3.dtd">
<nitf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
  <head>
    <title>
        Sorry, Ma'am, No Listing for 'enry 'iggins; Voice Recognition
        Is Improving, but Don't Stop the Elocution Lessons
    </title>
    <meta content="02ess" name="slug"/>
    <meta content="26" name="publication_day_of_month"/>
    <meta content="6" name="publication_month"/>
    <meta content="1995" name="publication_year"/>
    <meta content="Monday" name="publication_day_of_week"/>
    <meta content="Business/Financial Desk" name="dsk"/>
    <meta content="1" name="print_page_number"/>
    <meta content="D" name="print_section"/>
    <meta content="5" name="print_column"/>
    <meta content="Technology; Business" name="online_sections"/>
    <meta content="http://www.nytimes.com/1995/06/27/02ess.html" name="alternate_url"/>
    <meta content="Correction Appened" name="banner"/>
    <meta content="19950627T000000" name="correction_date"/>
    <meta content="EDUCATION" name="feature_page"/>
    <meta content="columnName" name="Education Column"/>
    <meta content="seriesName" name="Education Series"/>
    <docdata>
      <doc-id id-string="   "/>
      <doc.copyright holder="The New York Times" year="1995"/>
      <series series.name="Sorry, Ma'am, No Listing for 'enry 'iggins"/>
      <identified-content>
        <classifier class="indexing_service" type="biographical_categories">Books and Magazines</classifier>
        <classifier class="indexing_service" type="descriptor">DATA PROCESSING (COMPUTERS)</classifier>
        <location class="indexing_service">NEW YORK, NY</location>
        <classifier class="indexing_service" type="names">MCLEMORE, CYNTHIA</classifier>
        <org class="indexing_service">LINGUISTIC DATA CONSORTIUM</org>
        <person class="indexing_service">KAUFMAN, MICHAEL T</person>
        <object.title class="indexing_service">NEW YORK TIMES CORPUS (DATA)</object.title>
        <classifier class="online_producer" type="types_of_material">Article</classifier>
        <classifier class="online_producer" type="taxonomic_classifier">Top/News/Technology</classifier>
        <classifier class="online_producer" type="descriptor">Computers And The Internet</classifier>
        <classifier class="online_producer" type="general_descriptor">Research</classifier>
        <location class="online_producer">Philadelphia (Penna)</location>
        <org class="online_producer">Linguistic Data Consortium (LDC)</org>
        <person class="online_producer">Lomax, Alan</person>
        <object.title class="online_producer">New York Times Corpus (DATA)</object.title>
      </identified-content>
    </docdata>
    <pubdata date.publication="19950626T000000"
             ex-ref="http://query.nytimes.com/gst/fullpage.html?res=990CEFDC1139F935A15755C0A963958260"
             item-length="1590"
             name="The New York Times"
             unit-of-measure="word"
    />
  </head>
  <body>
    <body.head>
      <hedline>
        <hl1>Voice Recognition Is Improving, but Don't Stop the Elocution Lessons</hl1>
        <hl2 class="online_headline">Sorry, Ma'am, No Listing for 'enry 'iggins</hl2>
      </hedline>
      <byline class="print_byline">By MICHAEL T. KAUFMAN</byline>
      <byline class="normalized_byline">KAUFMAN, MICHAEL T</byline>
      <dateline>Philadelphia, June. 25</dateline>
      <abstract>
        <p>
            The Linguistic Data Consortium, a research cooperative,
            has released several large collections of data to spur advances
            in speech recognition.
        </p>
      </abstract>
    </body.head>
    <body.content>
      <block class="lead_paragraph">
        <p>What if I say "tomahto" and you say "tomayto?"</p>
      </block>
      <block class="online_lead_paragraph">
        <p>What if I say "tomahto" and you say "tomayto?"</p>
      </block>
      <block class="full_text">
        <p>As voice-recognition technologies are making their way from...</p>
      </block>
      <block class="correction_text">
        <p>Yesterday's article incorrectly stated...</p>
      </block>
    </body.content>
    <body.end>
      <tagline class="author_info">
        Michael T. Kaufman spent close to forty years at The New York
        Times as a reporter.
      </tagline>
    </body.end>
  </body>
</nitf>
```

Figure 1: Sample New York Times Corpus Document

2.1 Data Field Summary Table

Table 1 summarizes the data fields in the sample document presented above. The column values for this table are as follows.

1. **Short Name:** This column provides a short name for the data field referred to in the sample document. This naming convention allows for greater clarity in describing the corpus documents.
2. **Type:** The data type for the value in the specified field. Please note that this document defines the 'Integer' type as a 4 bytes integer and the 'Long' type as an 8-bit integer.
3. **Count:** The count column indicates if a document may contain only a single instance of the specified value or if it may contain multiple instances.
4. **XPATH:** The XPATH column provides an XPATH query that may be used to retrieve the specified data field from documents in the corpus. To learn more about XPATH, please refer to the w3c XPATH specification at http://www.w3.org/TR/xpath.
5. **Sample Path:** This column indicates the value of the specified data field in the sample document shown in Figure 1.

Short Name	Type	Count	XPATH	Sample Value
Alternate URL	URL	Single	/nitf/head/meta[@name="alternate_url"]/@content	http://www.nytimes.com/1995/06/27/02ess.html
Article Abstract	String	Single	/nitf/body/body.head/abstract	The Linguistic Data Consortium, a research cooperative…
Author Biography	String	Single	/nitf/body/body.content/block[@class="author_info"]	Michael T. Kaufman spent close to forty years at The New York Times…
Banner	String	Single	/nitf/head/meta[@name="banner"]/@content	Correction Appended
Biographical Categories	String	Multiple	/nitf/head/docdata/identified-content/classifier[@class="indexing_service" and @type="biographical_categories"]	Books and Magazines
Body	String	Single	/nitf/body/body.content/block[@class="full_text"]	As voice-recognition technologies are making their way from...
Byline	String	Single	/nitf/body/body.head/byline[@class="print_byline"]	By MICHAEL T. KAUFMAN
Column Name	String	Single	/nitf/head/meta[@name="column_name"]/@content	Education Column
Column Number	Integer	Single	/nitf/head/meta[@name="print_column"]/@content	5
Correction Date	Date	Single	/nitf/head/meta[@name="correction_date"]/@content	19950627T000000
Correction Text	String	Single	/nitf/body/body.content/block[@class="correction_text"]	Yesterday's article incorrectly stated...
Credit	String	Single	/nitf/head/docdata/doc.copyright/@holder	The New York Times
Dateline	String	Single	/nitf/body/body.head/dateline	Philadelphia, June. 25
Day Of Week	String	Single	/nitf/head/meta[@name="publication_day_of_month"]/@content	Monday
Descriptors	String	Multiple	/nitf/head/docdata/identified-	DATA PROCESSING

			content/classifier[@class="indexing_service" and @type="descriptor"]	(COMPUTERS)
Feature Page	String	Single	/nitf/head/meta[@name="feature_page"]/@content	EDUCATION
General Online Descriptors	String	Multiple	/nitf/head/docdata/identified-content/classifier[@class="online_producer" and @type="general_descriptor"]	Research
Guid	Long	Single	/nitf/head/docdata/doc-id/@id-string	771299
Headline	String	Single	/nitf/body[1]/body.head/hedline/hl1	Voice Recognition Is Improving, but Don't Stop the Elocution Lessons
Kicker	String	Single	/nitf/head/docdata/series/@series.name	Sorry, Ma'am, No Listing for 'enry 'iggins
Lead Paragraph	String	Single	/nitf/body/body.content/block[@class="lead_paragraph"]	What if I say "tomahto" and you say "tomayto?"
Locations	String	Multiple	/nitf/head/docdata/identified-content/location[@class="indexing_service"]	NEW YORK, NY
Names	String	Multiple	/nitf/head/docdata/identified-content/classifier[@class="indexing_service" and @type="names"]	MCLEMORE, CYNTHIA
News Desk	String	Single	/nitf/head/meta[@name="dsk"]/@content	Business/Financial Desk
Normalized Byline	String	Single	/nitf/body/body.head/byline[@class="normalized_byline"]	KAUFMAN, MICHAEL T
Online Descriptors	String	Multiple	/nitf/head/docdata/identified-content/classifier[@class="online_producer" and @type="descriptor"]	Computers And The Internet
Online Headline	String	Single	/nitf/body[1]/body.head/hedline/hl2	Sorry, Ma'am, No Listing for 'enry 'iggins
Online Lead Paragraph	String	Single	/nitf/body/body.content/block[@class="online_lead_paragraph"]	What if I say "tomahto" and you say "tomayto?"
Online Locations	Sring	Multiple	/nitf/head/docdata/identified-content/location[@class="online_producer"]	Philadelphia (Penna)
Online Organizations	String	Multiple	/nitf/head/docdata/identified-content/org[@class="online_producer"]	Linguistic Data Consortium (LDC)
Online People	String	Multiple	/nitf/head/docdata/identified-content/person[@class="online_producer"]	Lomax, Alan
Online Section	String	Single	/nitf/head/meta[@name="online_sections"]/@content	Business; Technology
Online Titles	String	Multiple	/nitf/head/docdata/identified-content/object.title[@class="online_producer"]	New York Times Corpus (DATA)
Organizations	String	Multiple	/nitf/head/docdata/identified-content/org[@class="indexing_service"]	Linguistic Data Consortium (LDC)
Page	Integer	Single	/nitf/head/meta[@name="print_page_number"]/@content	1
People	String	Multiple	/nitf/head/docdata/identified-content/person[@class="indexing_service"]	KAUFMAN, MICHAEL T
Publication Date	Date	Single	/nitf/head/pubdata/@date.publication	19950627T000000
Publication Day Of Month	Integer	Single	/nitf/head/meta[@name="publication_day_of_week"]/@content	26
Publication Month	Integer	Single	/nitf/head/meta[@name="publication_month"]/@content	06
Publication Year	Integer	Single	/nitf/head/meta[@name="publication_year"]/@content	1995
Section	String	Single	/nitf/head/meta[@name="print_section"]/@content	D
Series Name	String	Single	/nitf/head/meta[@name="series_name"]/@content	Education Series
Slug	String	Single	/nitf/head/meta[@name="slug"]/@content	02ess
Taxonomic Classifiers	String	Multiple	/nitf/head/docdata/identified-content/classifier[@class="online_producer" and	Top/News/Technology

			@type="taxinomic_classifier"]	
Titles	String	Multiple	/nitf/head/docdata/identified-content/object.title[@class="indexing_service"]	NEW YORK TIMES CORPUS (DATA)
Types Of Material	String	Multiple	/nitf/head/docdata/identified-content/classifier[@class="online_producer" and @type="types_of_material"]	Article
Url	URL	Single	/nitf/head/pubdata/@ex-ref	http://query.nytimes.com/gst/fullpage.html?res=990CEFDC1139F935A15755C0A963958260
Word Count	Integer	Single	/nitf/head/pubdata/@item-length	1590

Table 1: Data Field Overview

2.2 Data Field Details

This section provides detailed descriptions for the data fields summarized in Table 1.

2.2.1 Alternate Url

This field specifies the location on nytimes.com of the article. When present, this URL is preferred to the URL field on articles published on or after April 02, 2006, as the linked page will have richer content.

2.2.2 Author Biography

This field specifies the biography of the author of the article. Generally, this field is specified for guest authors not for New York Times reporters. When this field is specified for Times reporters, it is usually used to provide the author's email address.

2.2.3 Article Abstract

This field is a summary of the article written by the New York Times Indexing Service.

2.2.4 Banner

The banner field is used to indicate if there has been additional information appended to the articles since its publication. Examples of banners include ('Correction Appended' and 'Editor's Note Appended').

2.2.5 Biographical Categories

When present, the biographical category field generally indicates that a document focuses on a particular individual. The value of the field indicates the area or category in which this individual is best known. This field is most often defined for Obituaries and Book Reviews. These tags are hand-assigned by a team of library scientists working for the New York Times Indexing service.

Examples include:

1. Politics and Government (U.S.)
2. Books and Magazines
3. Royalty

2.2.6 Body

The body field is the text content of the article. Please note that this value includes the lead paragraph. Individual paragraphs for this field are enclosed in <p> tags.

2.2.7 Byline

This field specifies the byline of the article as it appeared in the print edition of the New York Times. Please note that not every article in this collection has a byline, as editorials and other types of articles are generally unsigned.

Sample bylines:

- By James Reston
- By JAMES GLANZ

2.2.8 Column Name

If the article is part of a regular column, this field specifies the name of that column.

Sample Column Names:

1. World News Briefs
2. WEDDINGS
3. The Accessories Channel

2.2.9 Column Number

This field specifies the column in which the article starts in the print paper. A typical printed page in the paper has six columns numbered from right to left. As a consequence most, but not all, of the values for this field fall in the range 1-6.

2.2.10 Correction Date

This field specifies the date on which a correction was made to the article. Generally, if the correction date is specified, the correction text will also be specified (and vice versa). This field is specified in the format YYYYMMDD'T'HHMMSS where:

1. YYYY is the four-digit year.

2. MM is the two-digit month [01-12].
3. DD is the two-digit day [01-31].
4. T is a constant value.
5. HH is the two-digit hour [00-23].
6. MM is the two-digit minute-past-the hour [00-59]
7. SS is the two-digit seconds-past-the-minute [00-59].

Please note that values for HH,MM, and SS are not defined for this corpus, that is to say HH,MM, and SS are always defined to be '00'.

2.2.11 Correction Text

For articles corrected following publication, this field specifies the correction. Generally, if the correction text is specified, the correction date will also be specified (and vice versa).

2.2.12 Credit

This field indicates the entity that produced the editorial content of this document. For this collection, the credit will always be set to 'The New York Times'.

2.2.13 Dateline

The 'dateline' field is the dateline of the article. Generally a dateline is the name of the geographic location from which the article was filed followed by a comma and the month and day of the filing.

Sample datelines:

- WASHINGTON, April 30
- RIYADH, Saudi Arabia, March 29
- ONTARIO, N.Y., Jan. 26

Please note:

1. The dateline location is the location from which the article was filed. Often times this location is related to the content of the article, but this is not guaranteed
2. The date specified for the dateline is often but not always the day previous to the publication date.
3. The date is usually but not always specified.

2.2.14 Day Of Week

This field specifies the day of week on which the article was published.

Must be one of:

- Monday
- Tuesday
- Wednesday
- Thursday
- Friday
- Saturday
- Sunday

2.2.15 Descriptors

The 'descriptors' field specifies a list of descriptive terms drawn from a normalized controlled vocabulary corresponding to subjects mentioned in the article. These tags are hand-assigned by a team of library scientists working in the New York Times Indexing service.

Examples Include:

- ECONOMIC CONDITIONS AND TRENDS
- AIRPLANES
- VIOLINS

2.2.16 Feature Page

This field specifies the name of the feature page on which the article appeared. A feature page is a themed page within a print section.

Examples Include:

- Consumer's World Page
- Society Desk
- Evening Hours Page

2.2.17 General Online Descriptors

The 'general online descriptors' field specifies a list of descriptors that are at a higher level of generality than the other tags associated with the article. These tags are algorithmically assigned and manually verified by nytimes.com production staff.

Examples Include:

- Surfing
- Venice Biennale
- Ranches

2.2.18 GUID

The GUID field specifies a (4-byte) integer that is guaranteed to be unique for every document in the corpus.

2.2.19 Headline

This field specifies the headline of the article as it appeared in the print edition of the New York Times.

2.2.20 Kicker

The kicker is an additional piece of information printed as an accompaniment to a news headline.

Examples Include:

- BASEBALL '87
- Bannu Journal
- BALKAN ACCORD
- Sports of The Times

2.2.21 Lead Paragraph

The 'lead Paragraph' field is the lead paragraph of the article. Generally this field is populated with the first two paragraphs from the article. Individual paragraphs for this field are enclosed in <p> tags.

2.2.22 Locations

The 'locations' field specifies a list of geographic descriptors drawn from a normalized controlled vocabulary that correspond to places mentioned in the article. These tags are hand-assigned by a team of library scientists working for the New York Times Indexing service.

Examples Include:

- Wellsboro (Pa)
- Kansas City (Kan)
- Park Slope (NYC)

2.2.23 Names

The 'names' field specifies a list of names mentioned in the article. These tags are hand-assigned by a team of library scientists working for the New York Times Indexing service.

Examples Include:

- Azza Fahmy

- George C. Izenour
- Chris Schenkel

2.2.24 News Desk

This field specifies the desk in the New York Times newsroom that produced the article. The desk is related to, but is not the same as the section in which the article appears.

2.2.25 Normalized Byline

The Normalized Byline field is the byline normalized to the form (last name, first name).

2.2.26 Online Descriptors

This field specifies a list of descriptors from a normalized controlled vocabulary that correspond to topics mentioned in the article. These tags are algorithmically assigned and manually verified by nytimes.com production staff.

Examples Include:

- Marriages
- Parks and Other Recreation Areas
- Cooking and Cookbooks

2.2.27 Online Headline

This field specifies the headline displayed with the article on nytimes.com. Often this differs from the headline used in print.

2.2.28 Online Lead Paragraph

This field specifies the lead paragraph as defined by the producers at nytimes.com. Individual paragraphs for this field are enclosed in <p> tags.

2.2.29 Online Locations

This field specifies a list of place names that correspond to geographic locations mentioned in the article. These tags are algorithmically assigned and manually verified by nytimes.com production staff.

Examples Include:

- Hollywood
- Los Angeles

- Arcadia

2.2.30 Online Organizations

This field specifies a list of organizations that correspond to organizations mentioned in the article. These tags are algorithmically assigned and manually verified by nytimes.com production staff.

Examples Include:

- Nintendo Company Limited
- Yeshiva University
- Rose Center

2.2.31 Online People

This field specifies a list of people that correspond to individuals mentioned in the article. These tags are algorithmically assigned and manually verified by nytimes.com production staff.

Examples Include:

- Lopez, Jennifer
- Joyce, James
- Robinson, Jackie

2.2.32 Online Section

This field specifies the section(s) on nytimes.com in which the article is placed. If the article is placed in multiple sections, this field will be specified as a ';' delineated list.

2.2.33 Online Titles

This field specifies a list of authored works mentioned in the article. These tags are algorithmically assigned and manually verified by nytimes.com production staff. Examples Include:

- Matchstick Men (Movie)
- Blades of Glory (Movie)
- Bridge & Tunnel

2.2.34 Organizations

This field specifies a list of organization names drawn from a normalized controlled vocabulary that correspond to organizations mentioned in the article. These tags are hand-assigned by a team of library scientists working in the New York Times Indexing service.

Examples Include:

- Circuit City Stores Inc
- Delaware County Community College (Pa)
- CONNECTICUT GRAND OPERA

2.2.35 Page

This field specifies the page of the section in the paper in which the article appears. This is not an absolute pagination. An article that appears on page 3 in section A occurs in the physical paper before an article that occurs on page 1 of section F.

2.2.36 People

This field specifies a list of people from a normalized controlled vocabulary that correspond to individuals mentioned in the article. These tags are hand-assigned by a team of library scientists working in the New York Times Indexing service. Examples Include:

- REAGAN, RONALD WILSON (PRES)
- BEGIN, MENACHEM (PRIME MIN)
- COLLINS, GLENN

2.2.37 Publication Date

This field specifies the date of the article's publication. This field is specified in the format YYYYMMDD'T'HHMMSS where:

1. YYYY is the four-digit year.
2. MM is the two-digit month [01-12].
3. DD is the two-digit day [01-31].
4. T is a constant value.
5. HH is the two-digit hour [00-23].
6. MM is the two-digit minute-past-the hour [00-59]
7. SS is the two-digit seconds-past-the-minute [00-59].

Please note that values for HH,MM, and SS are not defined for this corpus, that is to day HH,MM, and SS are always defined to be '00'.

2.2.38 Publication Day Of Month

This field specifies the day of the month on which the article was published, always in the range 1-31.

2.2.39 Publication Month

This field specifies the month on which the article was published in the range 1-12 where 1 is January 2 is February etc.

2.2.40 Publication Year

This field specifies the year in which the article was published. This value is in the range 1987-2007 for this collection.

2.2.41 Section

This field specifies the section of the paper in which the article appears. This is not the name of the section, but rather a letter or number that indicates the section.

2.2.42 Series Name

If the article is part of a regular series, this field specifies the name of that series.

2.2.43 Slug

The slug is a short string that uniquely identifies an article from all other articles published on the same day. Please note, however, that different articles on different days may have the same slug.

Examples of slugs include:

- 30other
- 12reunion

2.2.44 Taxonomic Classifiers

This field specifies a list of taxonomic classifiers that place this article into a hierarchy of articles. The individual terms of each taxonomic classifier are separated with the '/' character. These tags are algorithmically assigned and manually verified by nytimes.com production staff.

Examples Include:

- Top/Features/Travel/Guides/Destinations/North America/United States/Arizona
- Top/News/U.S./Rockies

- Top/Opinion

2.2.45 Titles

This field specifies a list of authored works that correspond to works mentioned in the article. These tags are hand-assigned by a team of library scientists working in the New York Times Indexing service.

Examples Include:

- Greystoke: The Legend of Tarzan, Lord of the Apes (Movie)
- Law & Order (TV Program)
- BATTLEFIELD EARTH (BOOK)

2.2.46 Types Of Material

This field specifies a normalized list of terms describing the general editorial category of the article. These tags are algorithmically assigned and manually verified by nytimes.com production staff.

Examples Include:

- REVIEW
- OBITUARY
- ANALYSIS

2.2.47 Url

This field specifies the location on nytimes.com of the article. The 'Alternative Url' field is preferred to this field on articles published on or after April 02, 2006, as the linked page will have richer content.

2.2.48 Word Count

This field specifies the number of words in the article, including the body, lead paragraph, headline and byline.

3. Production Process

Each article in The New York Times Corpus is the product of a production process that has evolved over the last two decades. Since the metadata included with the documents in this corpus reflects this process, it is important to provide an overview. The tile of each subsection indicates the time period in the corpus during which a particular phase of this process was carried out.

3.1 Content Creation (1987-2007)

In the content creation phase, the Author(s) of the article compose original copy.

3.2 Editing (1987-2007)

During the Editorial Production phase, individuals in the New York Time Newsroom refine the copy and decide where in the physical paper the article will appear. The article is then published.

3.3 Indexing (1987-2007)

In the indexing stage, library scientists at the New York Times Indexing Service incorporate the article into the New York Times Index. Annually published since 1913, the New York Times Index indexes the contents of the year's newspaper by grouping chronologically arranged summaries of articles under common subject headings. Figure 2 shows an example of the entry for Education and Schools from the 1993 Index.

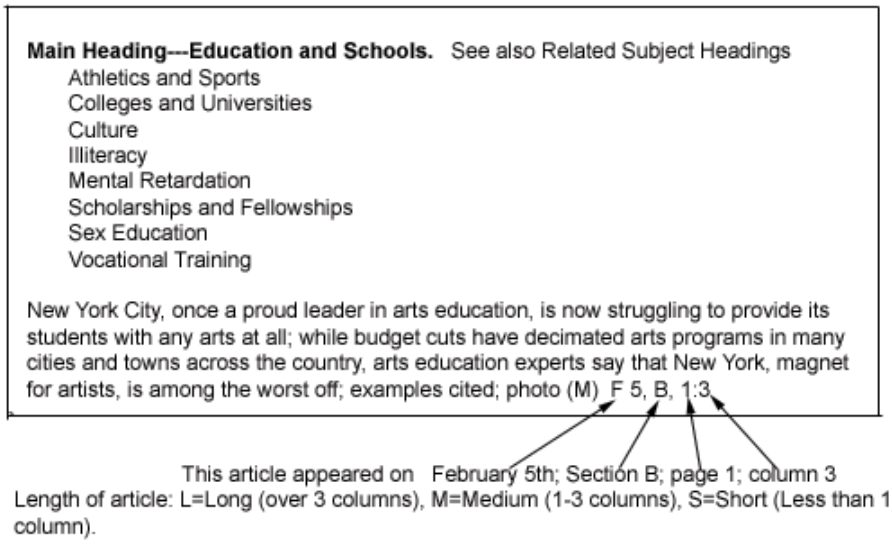

Figure 2: Sample Index Entry from 1993 Edition of The New York Times Index

To incorporate an article into the New York Times Index, the staff of the indexing service writes a brief summary of the article and tags the article with subject keywords drawn from a normalized controlled vocabulary. Although the indexing vocabulary slightly shifts over time (e.g. the "Greenhouse Effect" becomes "Global Warming"), an effort is made to ensure that articles are tagged in a consistent manner. For instance if one article makes mention of "Bill Clinton" and another refers to " President William Jefferson Clinton", they will both be tagged with "Clinton, Bill". As a further note, the terms in this indexing vocabulary are categorized into five groups: locations, organizations, people, subject descriptors, and titles of authored works (e.g. book titles). Examples of each of these categories are provided in section 2.2.

3.4 Online Production (2001-2007)

With the advent of the New York Times' online presence (www.nytimes.com), an online production phase was introduced to the newspaper's operations. In this phase, online producers augment articles with the data and metadata necessary to effectively present the content over the web.[1]

To place an article online, a producer starts by deciding which sections of the site the article should appear in (e.g. technology, fashion, etc). The producer might then create an online only headline for the article or opt to use the print headline.

Following that, the producer tags the article with relevant persons, places, organizations, titles and subject descriptors. As is the case with the indexing service, the tags used by the online production staff are drawn from a normalized controlled vocabulary and are applied in a consistent manner across articles. The tags used by the online producers are a subset of the tags used by the indexing service. Unlike the library scientists at the Times Indexing Service, online producers are assisted by an automated tagging algorithm that suggests potential tags for a given article. To ensure quality, producers review the suggested tags to add missing tags and remove irrelevant tags. Producers also use a similar approach to place the article into one or more nodes of a small taxonomy.

The online production process concludes with the publication of the article on nytimes.com.

3.5 Production Process Summary

Table lists the data fields in the corpus along with the phase of the production process during which each is produced.

Short Name	Source
Article Abstract	Indexing Service
Biographical Categories	Indexing Service
Descriptors	Indexing Service
Locations	Indexing Service
Names	Indexing Service
Organizations	Indexing Service
People	Indexing Service
Titles	Indexing Service
Word Count	Indexing Service
Author Biography	Newsroom
Banner	Newsroom
Body	Newsroom
Byline	Newsroom
Column Name	Newsroom

[1] Please note that this overview of the production process omits steps not relevant to the New York Times Corpus.

Column Number	Newsroom
Correction Date	Newsroom
Correction Text	Newsroom
Credit	Newsroom
Dateline	Newsroom
Day Of Week	Newsroom
Feature Page	Newsroom
Headline	Newsroom
Kicker	Newsroom
Lead Paragraph	Newsroom
News Desk	Newsroom
Normalized Byline	Newsroom
Page	Newsroom
Publication Date	Newsroom
Publication Day Of Month	Newsroom
Publication Month	Newsroom
Publication Year	Newsroom
Section	Newsroom
Series Name	Newsroom
Alternate Url	Online Production Staff
General Online Descriptors	Online Production Staff
Online Descriptors	Online Production Staff
Online Headline	Online Production Staff
Online Lead Paragraph	Online Production Staff
Online Locations	Online Production Staff
Online Organizations	Online Production Staff
Online People	Online Production Staff
Online Section	Online Production Staff
Online Titles	Online Production Staff
Slug	Online Production Staff
Taxonomic Classifiers	Online Production Staff
Types Of Material	Online Production Staff
Url	Online Production Staff

Table 2: Production Process Summary

4. Corpus Statistics

The New York Times Corpus contains 1,855,658 documents, covering a period of more than twenty years. Table 3 provides statistics about the distribution of the data fields over the corpus. The details of the columns of this table are as follows:

- **Short Name:** Same as in Table 1.
- **First Published On:** The publication date of the oldest article in the corpus containing the specified data field.
- **Documents Containing Field:** The number of documents in the corpus containing the specified data field.
- **Documents Containing Field (%):** The percentage of documents in the corpus containing the specified field.
- **Maximum Field Length:** The length of the largest value observed for fields of variable length. This field is left blank for numeric and date types, as they are not of variable length. Length is defined as follows for the following types:
  - **String:** The number of characters in the string.
  - **List:** The number of items in the list.
  - **URL:** The number of characters in the Url.
- **Total List Items:** For data fields that may appear more than once in a document, this column specifies the total number of instances of this data field in the corpus.

Short Name	First Published On	Documents Containing Field	Documents Containing Field (%)	Maximum Field Length	Total List Items
Alternate URL	8/1/00	328424	17.70%	97	
Article Abstract	9/3/04	664998	35.84%	2800	
Author Biography	9/23/90	52332	2.82%	6870	
Banner	1/1/87	42181	2.27%	45	
Biographical Categories	12/1/95	18041	0.97%	3	18745
Body	1/1/87	1831109	98.68%	232720	
Byline	1/1/87	1114053	60.04%	12755	
Column Name	6/12/96	40734	2.20%	115	
Column Number	1/1/87	1854136	99.92%		
Correction Date	1/1/87	41927	2.26%		
Correction Text	1/1/87	42178	2.27%	29941	
Credit	1/1/87	1855658	100.00%	18	
Dateline	1/1/87	439649	23.69%	14987	
Day Of Week	1/1/87	1855143	99.97%	9	
Descriptors	1/1/87	1574395	84.84%	42	3972094
Feature Page	10/22/90	16912	0.91%	3054	
General Online Descriptors	1/1/87	1479257	79.72%	44	4950495
Guid	1/1/87	1855658	100.00%		
Headline	1/1/87	1854654	99.95%	11276	
Kicker	1/1/87	561968	30.28%	159	
Lead Paragraph	1/1/87	1784878	96.19%	109935	
Locations	1/1/87	600114	32.34%	29	948625

Names	12/1/95	18418	0.99%	1	18418
News Desk	1/1/87	1855656	100.00%	78	
Normalized Byline	1/1/87	894105	48.18%	369	
Online Descriptors	9/17/00	281690	15.18%	213	589985
Online Headline	8/5/00	290021	15.63%	127	
Online Lead Paragraph	8/20/00	288753	15.56%	1989	
Online Locations	9/17/00	124174	6.69%	57	198484
Online Organizations	2/3/01	136993	7.38%	39	213992
Online People	2/7/01	114288	6.16%	41	179151
Online Section	1/1/87	1813489	97.73%	79	
Online Titles	6/7/01	7656	0.41%	5	7768
Organizations	1/1/87	596890	32.17%	40	901964
Page	1/1/87	1854516	99.94%		
People	1/1/87	1328045	71.57%	56	2372244
Publication Date	1/1/87	1855658	100.00%		
Publication Day Of Month	1/1/87	1855140	99.97%		
Publication Month	1/1/87	1855140	99.97%		
Publication Year	1/1/87	1855140	99.97%		
Section	1/1/87	1855638	100.00%	15	
Series Name	1/1/87	3599	0.19%	601	
Slug	3/15/87	987651	53.22%	40	
Taxonomic Classifiers	1/1/87	1846449	99.50%	103	8300051
Titles	1/1/87	166740	8.99%	18	194820
Types Of Material	1/1/87	770127	41.50%	147	796553
Url	1/1/87	1855658	100.00%	81	
Word Count	1/1/87	1855096	99.97%		

Table 3: Corpus Statistics