Metadata and Format Standardization Issues in Newspaper Databases and Publishing

Richard Geiger,
Library Director
San Francisco Chronicle
USA

Overview

• Early newspaper databases

• Newspaper database standardization

• Vendor database standardization

• Newspaper publishing format standards

Newspaper Databases - The Early Days

• The N.Y.Times Information Bank - 1971

• NYT online full-text capability - 1981.

• Louisville Courier Journal, INFO-KY - 1976

• Toronto Globe & Mail - 1977

• Philadelphia Inquirer and Daily News

• Vu/Text Information Services

• Knight-Ridder purchases Dialog Corp.

• Lexis-Nexis
Standardization By Newspapers

• NYT Information Bank - large subject list

• Others newspapers opted for a very short controlled vocabulary list

• Terms called “keywords”

• Knight-Ridder papers in Philadelphia, Detroit and Miami

Standardization By Vendors

• Newspaper library subjects unused

• Vender influence on document formats

• Inconsistencies

• QL software: only eight fields!

• New interest in newspaper database classification

• Lexis-Nexis has recently spent millions

Newspaper Document Formats

• Wire services or news distribution networks

• Associated Press, Reuters and Dow Jones

• 1970s: ITPC, International Press Telecommunications Council
Weaknesses of existing formats

- Need means of separating the meta-information from the body of the text
- Two examples:
- Need to identify names of companies and individuals
- Present worldwide date and time and automatically convert it into local date and time

Driving force: The Web

- Sever bonds to a particular business model or publishing technology
- Flexibility was imperative

The Solution

- A joint committee of the two primary standards organizations in the industry
- The NAA, the Newspaper Association of America
- The IPTC in Europe

NITF  (News Industry Text Format)

- The NITF is an XML-compliant markup language for news copy, press releases, wire services, newspapers, broadcasters, and Web-based news organizations.
- *Lingua franca*—for sharing and distributing files among traditional print publications systems, Web operations, archives and text resellers.

NITF in use

- Associated Press using NITF XML
- Reuters to use NewsML, another XML-based standard
NAA, IPTC and NITF

• The Newspaper Association of America

• http://www.naa.org/
  The IPTC web site has information on both formats

• http://www.iptc.org/

• For more information on NITF XML

  • http://www.nitf.org/

Some other IPTC projects

• Project codename IPTC2000 has been initiated to develop an XML based framework for structuring and managing Multimedia news

• Category Codes. A list of internationally acceptable subject names, subject matter names and some subject detail names for categorizing the content of News

Just for fun: Metadata in action

• SF Gate, site of The San Francisco Chronicle

  • http://www.sfgate.com/

• Archive search

  • http://www.sfgate.com/search/

• Restaurant Guide

  • http://www.sfgate.com/eguide/search/food/