

Lessons Learned in Content Architecture Harmonization and Metadata Models

Presented to
ISKO UK Conference
Shana Waggoner, Denise Bedford, Randi Park
June 22, 2009

World Bank publishing program

150 – 200 new titles/year in print and on-line

Three databases

World Bank e-Library

Global Economic Monitor Portal



Open Access website based on three World Bank Publications: www.dcp2.org



The screenshot shows a Windows Internet Explorer browser window displaying the DCP2 website. The address bar shows the URL <http://www.dcp2.org/main/Home.html>. The website header features the "DISEASE CONTROL PRIORITIES PROJECT" logo and a search bar. A navigation menu includes links for HOME, PUBLICATIONS, CUSTOM BOOKS AND SLIDES, NEWS AND EVENTS, TOOLS AND RESOURCES, BROWSE, and MY DCP2. The main content area is divided into several sections: a "Feature Story" about Leishmaniasis, "DCPP Publications" with a description of the second edition of "Disease Control Priorities in Developing Countries" and a "Priorities in Health" companion volume, an "About DCP2" section, and a "QUICK FIND" search tool. The taskbar at the bottom shows several open applications, including Microsoft Excel, World Bank Publications, and the DCP2 website itself. The system clock indicates it is 4:58 PM on Sunday, 6/7/2009.

DCPP - Home - Windows Internet Explorer
http://www.dcp2.org/main/Home.html
WB: People Directory provider

File Edit View Favorites Tools Help
Convert Select Links W3Schools WorldBank Intranet World Bank Publication Portal Interactive Factory
DCPP - Home

DISEASE CONTROL PRIORITIES PROJECT
SEARCH
GO
SEARCH TIPS

HOME | PUBLICATIONS | CUSTOM BOOKS AND SLIDES | NEWS AND EVENTS | TOOLS AND RESOURCES | BROWSE | MY DCP2

Feature Story:
Millions Still Suffer From Leishmaniasis

Spread via the bite of a sandfly, infections by leishmania parasites bring disfigurement, debility, disability, and death.

[Learn More](#) | [Past Features](#)

About DCP2

The Disease Control Priorities Project (DCPP) is an ongoing effort to assess disease control priorities and produce evidence-based analysis and resource materials to inform health policymaking in developing countries. DCPP has produced three volumes providing technical resources that can assist developing countries in improving their health systems and ultimately, the health of their people. [More](#)

DCPP Publications

Disease Control Priorities in Developing Countries
The second edition of this seminal work (DCP2), written by more than 350 specialists in diverse fields from around the world, provides the results of in-depth research, offers insightful analyses, and proposes context-sensitive policy recommendations to significantly reduce the burden of disease in developing countries and to improve the quality of life for all people.

Priorities in Health
This companion volume distills the essence of DCP2 into a succinct and readable format, providing

Browse by:
[Disease/Condition](#)
[Region/Country](#)
[Topic](#)
[Intervention](#)

Download Data Files

Create and download your custom book

QUICK FIND

Find specific content in the three core DCP2 publications.

Region / Country
Disease / Condition
Topic

Go

Internet | Protected Mode: On | 100%

Baker and Taylor ... Microsoft Excel World Bank Publi... Extend360 Mail - Inbox - IB... ISKO preesentatio...
New Bulletin Boa... DOL.pptx presentation.ppt ... Welcome to Worl... DCP2 - Home - ...

4:58 PM
Sunday
6/7/2009

Online subscription products World Bank e-Library

World Bank home - Windows Internet Explorer

http://www.worldbank.icebox.ingenta.com/?jsessionid=bl95ps35uej7g

WB: People Directory provider

File Edit View Favorites Tools Help

Convert Select Links W3Schools WorldBank Intranet World Bank Publication Portal Interactive Factory

DCPP - Home World Bank home

The World Bank
e LIBRARY

About the E-Library How to Subscribe New Content Alerts For Librarians Useful Links Contact

Adjust font size: A A A

WELCOME TO THE NEW WORLD BANK E-LIBRARY

The World Bank e-library is an online, fully cross-searchable portal of over 5,000 World Bank documents. The collection consists of over 2,000 World Bank publications and over 3,200 Policy Research Working Papers, plus each new book and paper as they are published. As an e-Library subscriber you gain unlimited access to a wealth of knowledge and expertise on development issues throughout the years and across a wide range of subjects.

CONDITIONAL CASH TRANSFERS: REDUCING PRESENT AND FUTURE POVERTY
BY ARIEL FISZBEIN , NORBERT SCHADY

Conditional Cash Transfer (CCT) programs aim to reduce poverty by making welfare programs conditional upon the receivers' actions.

[read more](#)

SEARCH Site Help

Go

Advanced Search

BROWSE BY REGION
Please Choose

BROWSE BY SUBJECT
Please Choose

BROWSE BOOKS
BROWSE WORKING PAPERS
BROWSE ALL JOURNALS

Signed in as:
+ WorldBank ePortal Trial

Sign out
Additional sign in

Done

Internet | Protected Mode: On 100%

Baker and Taylor ... Microsoft Excel World Bank Publi... Extend360 Mail - Inbox - IB... ISKO presentatio...

New Bulletin Boa... DOI.pptx presentation.ppt ... Welcome to Wor... World Bank hom...

5:27 PM
Sunday
6/7/2009

World Development Indicators

The screenshot shows a Windows Internet Explorer browser window displaying the 'DDP Quick Query' application. The address bar shows the URL: <http://ddp-ext.worldbank.org/ext/DDPQQ/member.do?method=getMembers&userid=1&queryId=135>. The page header includes 'The World Bank Group' logo and the text 'Quick Query selected from World Development Indicators'. A language dropdown menu is set to 'English'. Below the header, there are tabs for 'Select Variables' and 'Show Report', with 'Info' and 'Help?' links. A navigation bar contains 'Country', 'Series', and 'Time'. The main content area is titled 'Classification' and features a search box with 'Search', 'Reset', and 'Notes' buttons. A list of 'Available - 227' countries is shown in a scrollable box, including Afghanistan, Albania, Algeria, American Samoa, Andorra, Angola, Antigua and Barbuda, Argentina, and Armenia. Below this list are 'Select', 'Select All', 'Deselect', and 'Deselect All' buttons. A 'Selected - 0' section is currently empty, with a 'Next >>' button to its right. The Windows taskbar at the bottom shows several open applications, including Microsoft Excel, World Bank P..., Extend360, Mail - Inbox..., ISKO present..., New Bulletin..., and DDP Quick Q... The system tray indicates the time is 5:32 PM on Sunday, 6/7/2009.

Global Development Finance

The screenshot shows the DDP Quick Query web application running in Internet Explorer. The browser's address bar displays the URL: <http://ddp-est.worldbank.org/ext/DDPQQ/member.do?method=getMembers&userid=1&queryId=136>. The page header includes "The World Bank Group Quick Query GDF Online Limited trial version" and a language selector set to "English".

The main interface is divided into several sections:

- Select Variables:** Includes tabs for "Country", "Series", and "Time".
- Classification:** A list of regions with checkboxes: Aggregates, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, South Asia, and Sub-Saharan Africa.
- Search:** A search box with "Search", "Reset", and "Notes" buttons.
- Available - 7:** A list of countries: All developing countries, Burkina Faso, Costa Rica, Nepal, Tajikistan, Togo, and Yemen, Rep.
- Selected - 0:** A list for selected items, currently empty.

Buttons for "Select", "Select All", "Deselect", and "Deselect All" are located between the available and selected lists. A "Next >>" button is at the bottom right of the selection area.

At the bottom of the page, there is a copyright notice: "© 2007 The World Bank Group, All Rights Reserved."

The Windows taskbar at the bottom shows the system tray with the date "Tuesday 6/9/2009" and time "10:15 AM". Several application windows are open, including "DDP Quick Query", "ISKO slides from S...", "isko-presentation.p...", and "presentation.ppt [C...".

Africa Development Indicators

The screenshot displays the 'Quick Query' interface for 'Africa Development Indicators (ADI) Data Availability'. The browser window title is 'DDP Quick Query - Windows Internet Explorer'. The address bar shows the URL: <http://ddp-ext.worldbank.org/ext/DDPQQ/member.do?method=getMembers&userid=1&queryId=229>. The page header includes 'The World Bank Group' and 'Quick Query Africa Development Indicators (ADI) Data Availability'. A language dropdown is set to 'English'. Below the header, there are tabs for 'Select Variables' and 'Show Report', along with 'Info' and 'Help?' links. The 'Country' section is active, showing a search box and a list of available countries: Africa, Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, and Cape Verde. The 'Selected' list is currently empty. The footer of the page contains the copyright notice: '© 2007 The World Bank Group. All Rights Reserved.' The Windows taskbar at the bottom shows the system tray with the date 'Tuesday 6/9/2009' and time '10:16 AM'. The taskbar also displays several open applications, including 'DDP Quick Query - ...', 'ISKO slides from S...', 'isko-presentation.p...', and 'presentation.ppt [C...'.

Global Economic Monitor

Global Economic Monitor Intro - Windows Internet Explorer

http://siteresources.worldbank.org/EXT/GBLPROSPECTS/APRIL/Resources/659148-1178128947216/GEMhome.html

File Edit View Favorites Tools Help

Convert Select Links W3Schools WorldBank Intranet World Bank Publication Portal Interactive Factory

Global Economic Monitor Intro

Global Economic Monitor

Home > Data & Research > Prospects > Global Economic Monitor Intro

- GEM Intro
- GEM Login
- Subscription Info
- Monthly Briefs
- Indicators
- Regions
- Data Access
- Forecasts
- Publications
- Site Tools
- Contact Us

Global Economic Monitor Intro

GEM is a new World Bank online subscription service, updated daily, providing analysis of current economic trends, economic and financial data indicators, and an extensive database.

Focus Topics

- Global food prices jumped 33%, largely because of increased production of bio-fuels... [login/subscribe](#)**
25 January 2008
- Commodity price impacts. A measure of the real oil price increased 39% in the year to November... [login/subscribe](#)**
12 December 2007
- Terms of trade loss made worse for some oil-importing, low-income countries... [login/subscribe](#)**
10 November 2007

Monthly Briefs

Global Trends

Publications

Commodity Markets Review

Macroeconomic Trends - Monthly

Scroll-over to select: Domestic External

Domestic Indicators

Country: Albania

	2007	YTD08	MRV*
CPI, %	2.9	3.4	2.6
M2/reserves	4.2	4.4	4.0
M2-multiplier	3.3	3.5	3.3
Real deposit rate	2.7	3.2	4.1

varies. Period averages. Sources: World Bank, Datastream and IMF. [Notes](#) [login](#)

Related World Bank Products

Global Development Finance 2007

Daily Brief

The Daily Brief is compiled each weekday morning to provide an update of the latest developments in world macro-, financial- and commodity markets. [login](#)

Daily Chart

European retail sales down 1% in February (y/y) on pickup in Germany
retail sales: Euro Area, Germany, France and Italy, c/y 3mna y/y

Source: Eurostat.

Commodity Market Trends - Daily

Scroll-over to select: Energy Non-Energy

Energy Markets

Series: Avg oil (Brent/Dub/WTI) \$/B

	2008	YTD	MRV*
World	97.0	48.1	67.1

*MRV=Most recent (close of business) daily value. Period averages. Sources: World Bank and Datastream. [Notes](#) [login](#)

Financial Market Trends - Daily

Country: Argentina

	2009	YTD	MRV*
LCU/USD		3.61	3.74
Policy rate		12.24	10.81
Spread, bp		1585	1158
EMBI+		52.0	49.5
Stk ind. %LCU		-54.4	-53.1
Stk ind. %US\$		-60.4	-61.3

*MRV=Most recent (cob) daily value. Sources: World Bank and Datastream. Availability varies. [Notes](#) [login](#)

Done, but with errors on page.

Unknown Zone (Mixed) | Protected Mode: Off | 100%

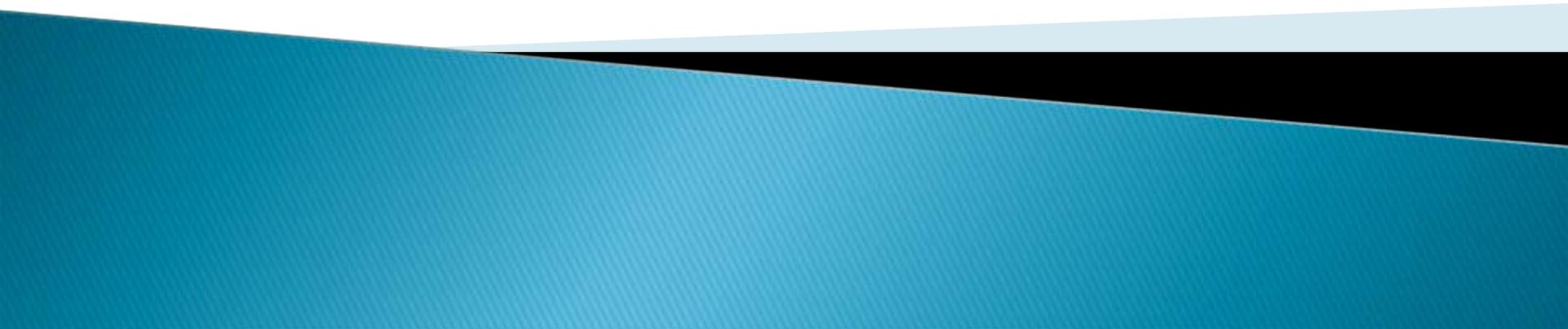
Global Economic M... ISKO slides from S... isko-presentation.p... presentation.ppt [C...

Tuesday 6/9/2009 10:19 AM

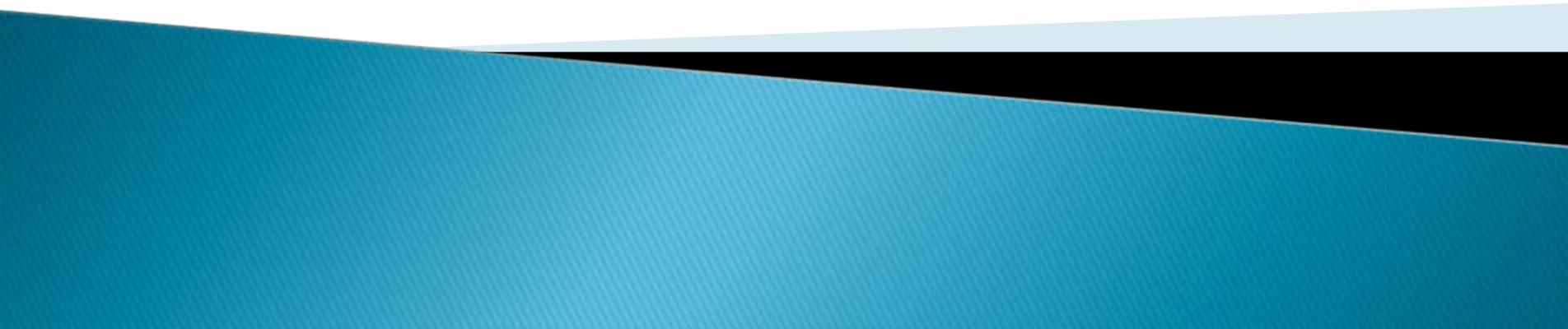
Identifying Broad Goals

- ▶ Mission–critical goals
 - Raising awareness of poverty
 - Underscoring need for economic and social development
 - Disseminating Bank knowledge, programs, and involvement
 - Thinking ahead in terms of overall program

Why the World Development Report?

- ▶ World Bank flagship
 - ▶ Most accessible publication
 - ▶ Broadest audience
 - ▶ Legacy content of high value
 - ▶ Need to preserve in viable, neutral, reusable format
- 

The WDR Project Must . . .

- ▶ Be replicable
 - ▶ Design architecture and tools with consistency and evolution in mind
 - ▶ Migrate from DVD to online presence
 - ▶ Build on metadata models and semantics for future use and integration
 - ▶ Address needs of partners like crossref.org
 - ▶ Suit broader base of users with data mining etc
- 

Analyzing the Series 1

- ▶ Evolution of content over 30 years
 - Author team, topic, focus
 - World Bank background, introduction to global economics/finance, and historical issues
 - Balance of historical coverage, topical coverage, data
 - Role and normalization of data
- ▶ ISBNs, DOIs, and other backbone numbers
- ▶ Format and availability over 30 years

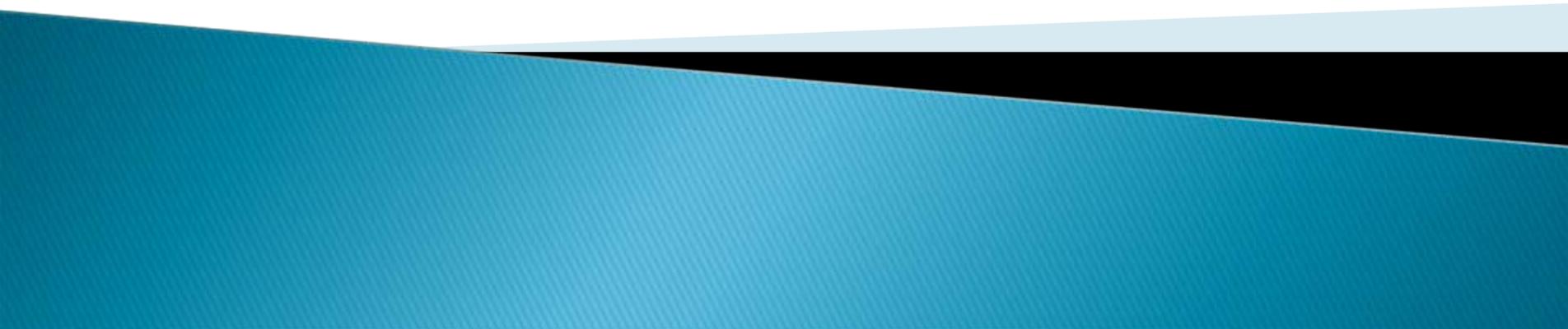
Analyzing the Series 2

- ▶ Assessing features and structure
 - Different relations between parts and chapters
 - Varying roles of Overviews and Introductions
 - Special features
 - Cross reference styles
 - Citations, reference lists, and endnotes
 - Nomenclature
- ▶ Much documentation needed for XML conversion
 - Some lessons

Early Content Decisions

- ▶ Include entirety of each Report?
 - XML vs PDF
 - Front matter and back matter
 - Data
 - Tables, figures, boxes, equations
- ▶ Create, integrate, or plan for additional content?
 - Abstracts
 - Background papers
 - World Bank–specific information

Overarching Concerns

- ▶ What do researchers expect from World Bank content?
 - ▶ What is appropriate for WDR content pool?
 - ▶ What tools and functions could be modified for future uses and new content?
 - ▶ What do we need to learn about our metadata model?
 - ▶ Shhh: Timeframe, budget
- 



the complete world development report 1978-2009

GO

[Advanced Search](#)[HOME](#)[BROWSE BY
TITLE](#)[BROWSE BY
TOPIC](#)[DEVELOPMENT
DATABASE](#)[REGIONS](#)[TIMELINE](#)[ABOUT](#)

The World Bank's *World Development Report*, published annually since 1978, is an invaluable guide to the economic, social, and environmental state of the world today. Each report provides in-depth analysis and policy recommendations on a specific and important aspect of development—from agriculture, the role of the state, transition economies, and labor to infrastructure, health, the environment, and poverty.

Through the quality and timeliness of the information it provides, the report has become a highly influential publication that is used by many multilateral and bilateral international organizations, national governments, scholars, civil society networks and groups, and other global thought leaders to support their decision-making processes.

To commemorate the *World Development Report's* 30th anniversary, the World Bank has brought together the full collection of reports in an easy-to-use, fully searchable digital archive: *The Complete World Development Report*. This reference tool allows users to quickly and easily search both within and across reports as well as to browse reports by title or topic. Additionally, through the up-to-date database derived from *World Development Indicators*, users can access up to 30 years of statistical data for 50 development indicators for 210 countries and 18 country groups.

We hope that this collection will help build a better understanding of today's most pressing development issues and increase awareness of the world's most critical challenge: putting an end to global poverty.

World Development Report

Explore the full texts of all reports, organized by title.

Development Database

Access more than 50 development indicators with time series data for more than 200 countries for up to 30 years.

WDR 2009: Reshaping Economic Geography

See the most recent report, which examines the interactions between economic geography, growth, and living standards.

Browsing Strategies

- ▶ Browse to what?

- ▶ Principles

- Title
- Chronology
- Topic
- Region
- Country

- ▶



the complete world development report 1978-2009

type search term(s) here.

GO

[Advanced Search](#)



[HOME](#)

[BROWSE BY TITLE](#)

[BROWSE BY TOPIC](#)

[DEVELOPMENT DATABASE](#)

[REGIONS](#)

[TIMELINE](#)

[ABOUT](#)

Browse by Topic

Browse World Development Reports by topic. Click a topic or + for a list of reports on that topic. Click a title to go to that report.

Agriculture and Rural Development

SORT BY : [Oldest](#) | [Newest](#)

[WDR 2008 : Agriculture for Development](#)

[WDR 1986 : Trade and Pricing Policies in World Agriculture](#)

[WDR 1982 : Agriculture and Economic Development](#)

Banking, Finance, and Investment

Education and Training

SORT BY : [Oldest](#) | [Newest](#)

[WDR 2007 : Development and the Next Generation](#)

[WDR 2004 : Making Services Work for Poor People](#)

[WDR 1998/99 : Knowledge for Development](#)

Energy, Industry, and Mining

Environment and Pollution Prevention

Labor and Income

Legal and Judicial Issues

Poverty and Poverty Reduction

SORT BY : [Oldest](#) | [Newest](#)

[WDR 2006 : Equity and Development](#)

[WDR 2000/2001 : Attacking Poverty](#)

[WDR 1990 : Poverty](#)

[WDR 1980 : Poverty and Human Development](#)

[WDR 1978 : Prospects for Growth and Alleviation of Poverty](#)

Private Sector

Public Policy

Social and Cultural Issues

Technology and Telecommunications

the complete world development report 1978-2009

GO

[Advanced Search](#)


HOME

BROWSE BY
TITLEBROWSE BY
TOPICDEVELOPMENT
DATABASE

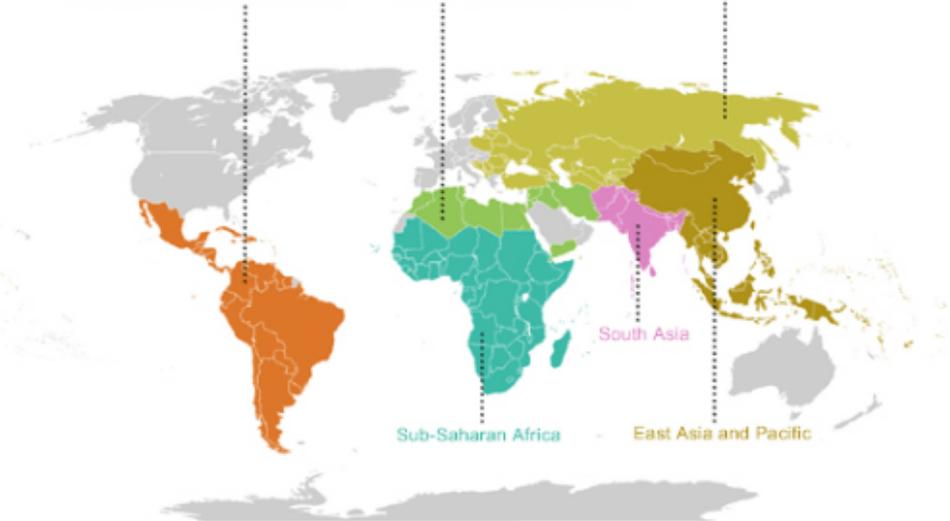
REGIONS

TIMELINE

ABOUT

Regions

Latin America and the Caribbean Middle East and North Africa Europe and Central Asia



The regions used in this DVD are based on those created and defined by the World Bank for analytical and operational purposes. These regions may differ from common geographic usage or from the regions defined by other organizations.

Regional groupings and the aggregate measures for regions include only low- and middle-income economies. Data below are for 2006.

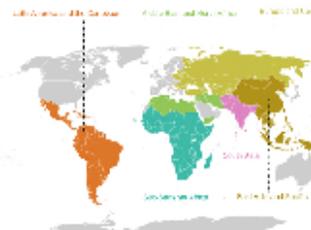
Africa (Sub-Saharan)

This region is home to 781.8 million people and has a population growth rate of 2.6 percent, the highest of all the Bank regions. Average life expectancy is 50 years, the same as in 1990. The region's GNI per capita is \$920. This region includes the following countries and

[Regions](#)
[Income Levels](#)

Development Database

Access more than 50 development indicators with series data for more than 200 countries for up to 30 years.



Search Strategies

- ▶ Search all content?
- ▶ Search to what?
- ▶ HTML vs PDF?
- ▶ Metadata
- ▶ Search fields
- ▶ Boolean logic
- ▶ Data mining
- ▶ Navigation



the complete world development report 1978-2009

type search term(s) here.

GO

Advanced Search



HOME

BROWSE BY TITLE

BROWSE BY TOPIC

DEVELOPMENT DATABASE

REGIONS

TIMELINE

ABOUT

Advanced Search

To perform a search, use any combination of search term(s) in the text field and titles, topics, and regions. The default is to search all titles, topics, and regions

Development Database

Access more than 50 development indicators and time series data for more than 200 countries to 30 years.

Search Tips

type search term(s) here.

Search

Reset all fields

Use AND, OR, NOT, or quotation marks.

Titles [Select All](#) | [Deselect All](#)

- WDR 2009:Reshaping Economic Geography
- WDR 2008:Agriculture for Development
- WDR 2007:Development and the Next Generation
- WDR 2006:Equity and Development
- WDR 2005:A Better Investment Climate for Everyone
- WDR 2004:Making Services Work for Poor People
- WDR 2003:Sustainable Development in a Dynamic World: ...
- WDR 2002:Building Institutions for Markets
- WDR 2000/2001:Attacking Poverty

Topics [Select All](#) | [Deselect All](#)

- Agriculture and Rural Development
- Banking, Finance, and Investment
- Education and Training
- Energy, Industry, and Mining
- Environment and Pollution Prevention
- Gender
- Governance, Civil Society, and Participation
- Health, Nutrition, and Population
- Infrastructure, Transport, and Urban Development



the complete world development report 1978-2009

type search term(s) here.

GO

[Advanced Search](#)



[HOME](#)

[BROWSE BY TITLE](#)

[BROWSE BY TOPIC](#)

[DEVELOPMENT DATABASE](#)

[REGIONS](#)

[TIMELINE](#)

[ABOUT](#)

Search Results

44 items for youth

[Modify Search](#)

[New](#)

Page 1 of 5

Showing 1 - 10 of 44 items

[1](#) [2](#) [3](#) [4](#) [5](#) [Next>](#)

Go to page:

[View All \(44\)](#)

Sort by: [Relevance](#)

1. Spotlight. Do baby booms lead to employment busts? Not in OECD countries
WDR 2007-Development and the Next Generation

[The post...]

2. Chapter 9: Youth policy: Doing it and getting it right
WDR 2007-Development and the Next Generation

[In a world that demands more than basic skills, expanding the opportunities for young people to learn means going beyond primary schooling. Because building human capital takes more than schooling, it also means including young people more in their communities and societies so that they have opportunities to use their talents at work and to

3. Spotlight. What donors can do
WDR 2007-Development and the Next Generation

[Donors can help governments undertake policies to broaden opportunities, enhance capabilities, and provide second chances to young people. Key actions for donors: applying the youth lenses of opportunities, capability, and second

Narrow your Result

Click a region or topic below to narrow your results. Click 'view all' to return to all results.

[View All](#)

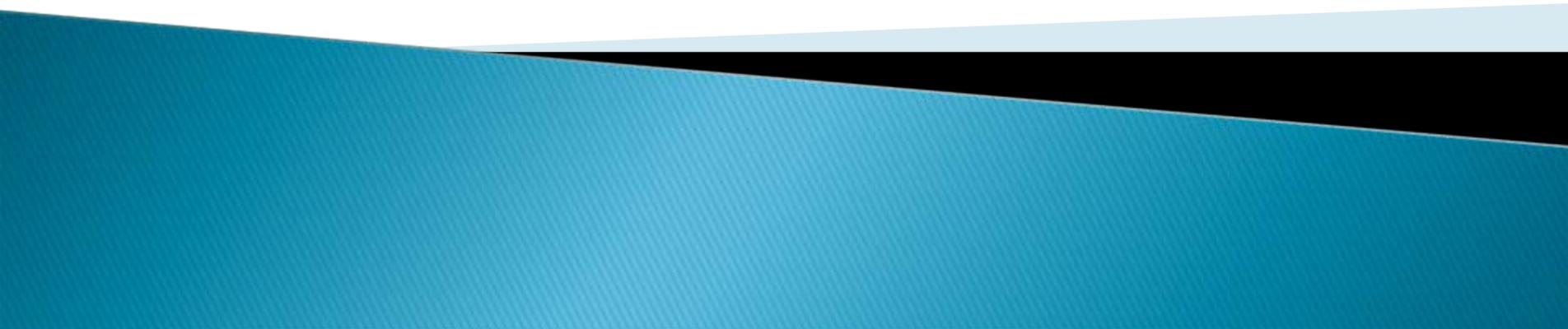
Regions

- [East Asia & the Pacific](#)
- [Europe & Central Asia](#)
- [Latin America and the Caribbean](#)
- [Middle East and North Africa](#)
- [South Asia](#)
- [Sub-Saharan Africa](#)

Topics

- [Agriculture and Rural Development](#)
- [Banking, Finance, and Investment](#)
- [Education and Training](#)
- [Energy, Industry, and Mining](#)
- [Environment and Pollution Prevention](#)

Presentation and Linking

- ▶ Overall architecture
 - ▶ Navigable Tables of Contents
 - ▶ Endnotes and references
 - ▶ Tables, figures, and boxes
 - ▶ Search-based navigation
 - ▶ Lessons learned
- 



the complete world development report 1978-2009

type search term(s) here.

GO

Advanced Search



HOME

BROWSE BY TITLE

BROWSE BY TOPIC

DEVELOPMENT DATABASE

REGIONS

TIMELINE

ABOUT

Home > Search Results: "youth" > WDR 2007 : Chapter 9: Youth policy: Doing it and getting it right

< Back to Full Search Results | Previous Match | Next Match >

Search this book | PDF of the book | Print

< Back Forward >

FOOTNOTES

Chapter 9 - Youth policy: Doing it and getting it right

Introduction

In a world that demands more than basic skills, expanding the opportunities for young people to learn means going beyond primary schooling. Because building human capital takes more than schooling, it also means including young people more in their communities and societies so that they have opportunities to use their talents at work and to participate as active citizens. Informing and facilitating their decision making—so that they choose well—adds to their success as workers and entrepreneurs, as parents, and as citizens. And for those who have to recover from poor decisions or poor circumstances providing second chances—to make up for missed opportunities—can keep young people from being irrevocably left behind.

For many low-income countries, the priority is to provide quality basic education (including lower secondary education) for young adolescents and, where HIV/AIDS prevalence is high, such as in parts of Sub-Saharan Africa, to safeguard health by preventing transmission among young people who are beginning to be sexually active. Both will help in the transition to work. For many young people in these countries, basic skills for life and work will have to be acquired through second-chance programs. Making full use of young people's skills and providing them further opportunities to build skills on the job will require maintaining an overall macroeconomic environment conducive to growth.

For middle-income and high-growth countries that have already provided mass basic and then secondary

Book Details



WDR 2007 : Development and the Next Generation

Author : The World Bank

Publisher : The World Bank

Pub. Date : 2006

ISBN : 978-0-8213-6541-0

Table of Contents

EXPAND : ALL

COLLAPSE : ALL

+ Abstract

+ Front Matter

Knowledge Architecture

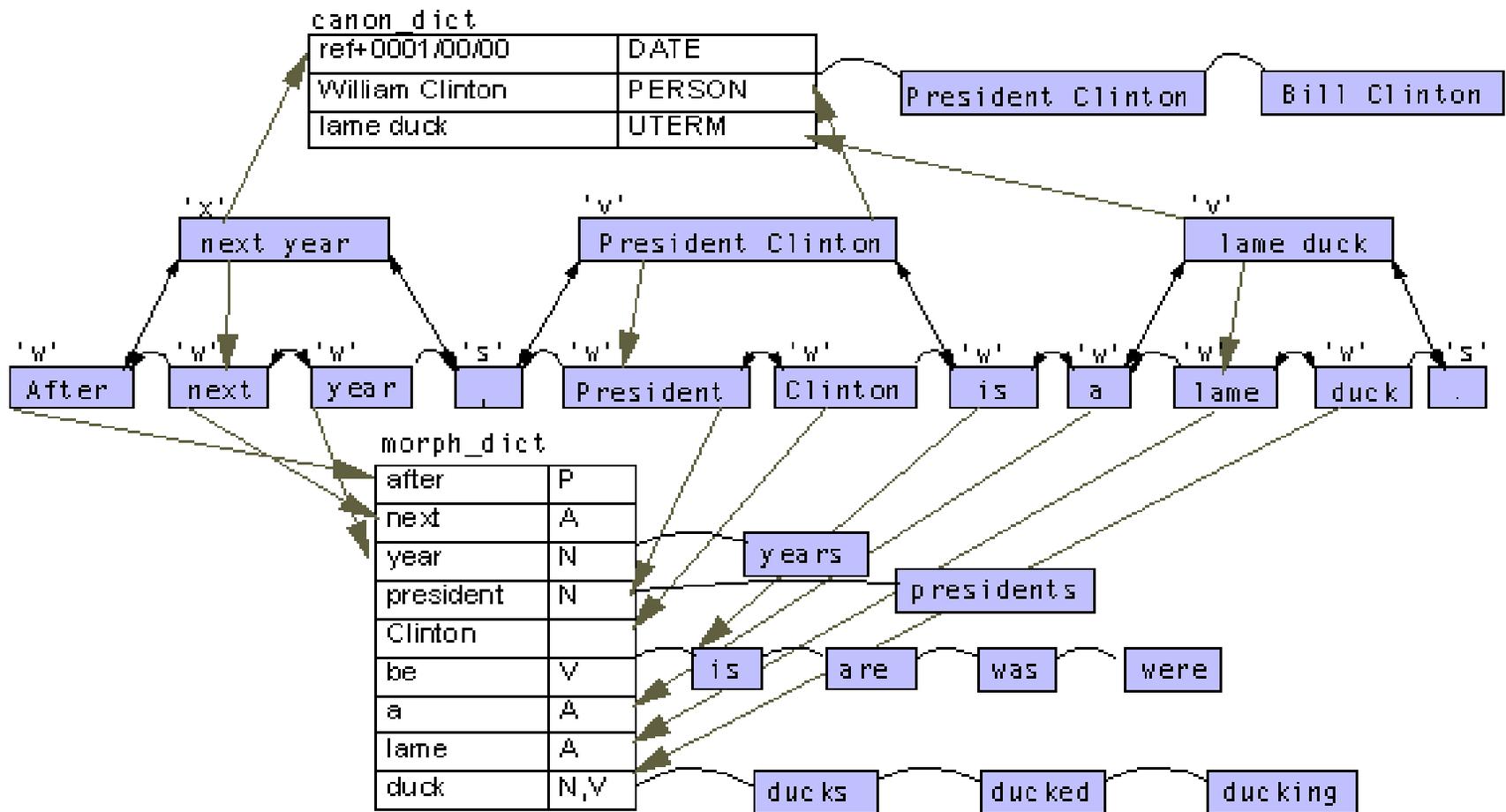
- ▶ Semantic analysis technologies are used to generate persistent metadata for core attributes
 - ▶ Engines use Natural Language Processing technologies to create semantic index of document
 - ▶ Reference Sources (taxonomies) are built into the profiles
 - ▶ Work with business units to build profiles (knowledge engineering)
- 

What is Natural Language Processing?

- ▶ **Natural language processing**
 - has been around for a long time – since the 1950s when we first had the computing capacity to explore language behavior
 - is a subfield of artificial intelligence and linguistics. It studies the problems of automated generation and understanding of natural human languages.
 - encompasses a wide range of approaches, including -- Information extraction, Automatic Summarization, Natural language generation, Text to speech conversion, Speech recognition, Machine translation, Question answering, Information retrieval, Text-proofing, Translation technology

- ▶ **Natural language understanding systems** convert samples of human language into more formal representations that are easier for computer programs to manipulate.

Example of Semantic Analysis



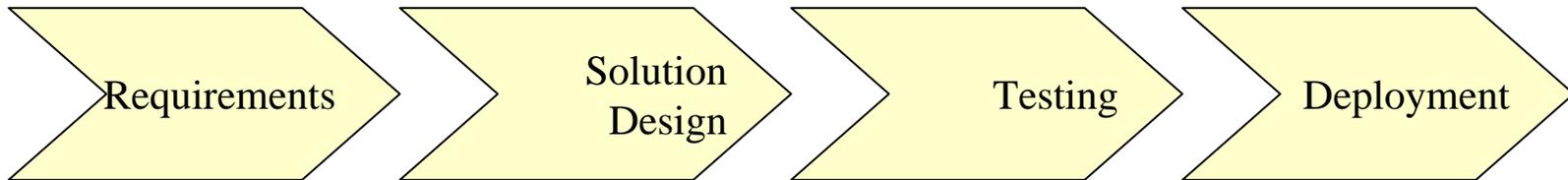
How the Technology Works

- ▶ Detects language (dictionary, grammars, etc.)
- ▶ Calls language rules and knowledgebase
- ▶ Generates semantic index of the content
- ▶ Takes instructions re: what to do with the indexed document
- ▶ Interprets the profile instructions to:
 - Apply a concept extraction profile
 - Apply a categorization profile
 - Apply summarization rules
 - Ingest a list of terms and applying rules for clustering
 - To decide how to output the results

Knowledge Engineering

- ▶ Simulates human decision making and thinking
 - ▶ We embed in Teragram the knowledge base that a person would use to make a decision
 - ▶ We build the rules to represent how a person would make a decision
 - ▶ Goal is to achieve 90% equivalence to human decisions at the prototype, and continue to refine until we reach 100% or greater
- 

How Knowledge Engineering Works



- Problem Definition?
- How will be used?
- What content?
- Expected result?
- Existing human Process?
- How done now?
- How does person Do It?
- One-time or on-going?

- Model thinking or Decision process
- Define rules
- Identify knowledge Sources
- Pre-processing?
- Post-processing?
- How to access content?
- How to integrate with applications?

- Build profile Or rule set
 - Test on training set
 - Review with business
 - Feedback
 - Refine
 - Business signoff
- Develop API's
Define output formats

- Operationalize profile
- Operationalize Process ing
- Provide training to business
- Establish ongoing feedback

Building the Profiles

- ▶ Each metadata attribute has its own semantic profile
- ▶ Profile consists of rules and patterns that simulate how a person would make a decision or derive the metadata value – for example:
 - Topic, Country, Content Type, Records Class all follow human “classification” processes
 - Author, Keywording, Series Name, Publisher all follow human concept extraction process
- ▶ Each profile has its own distinct set of rules

Automatically Generated XML Metadata

```
<?xml version="1.0" encoding="utf-8" standalone="yes" ?>
- <TERAGRAM_XML>
- <RESPONSE>
  - <ENVIRONMENT>
    <SOURCE name="W:\Concept Extraction\FSAP Training Set\Disc2\<input type="text" value="" /> type="file" />
    <CATEGORY_SET name="ImageBank-EN" />
  </ENVIRONMENT>
  <LANGUAGE codepage="CP1252" name="ENGLISH" />
- <CATEGORY name="Topics">
  <CATEGORY_ITEM mdk="672884" name="Banks and Banking Reform" relevant="1240.0" />
  <CATEGORY_ITEM mdk="672893" name="Financial Intermediation" relevant="698.0" />
  <CATEGORY_ITEM mdk="672634" name="Economic Theory and Research" relevant="552.0" />
  <CATEGORY_ITEM mdk="672888" name="Financial Crisis Management and Restructuring" relevant="551.0" />
  <CATEGORY_ITEM mdk="672690" name="Banking Law" relevant="446.0" />
  <CATEGORY_ITEM mdk="672685" name="Settlement of Investment Disputes" relevant="439.0" />
  <CATEGORY_ITEM mdk="757896" name="Corporate Law" relevant="413.0" />
  <CATEGORY_ITEM mdk="672642" name="Investment and Investment Climate" relevant="400.0" />
  <CATEGORY_ITEM mdk="672885" name="Insurance and Risk Mitigation" relevant="394.0" />
  <CATEGORY_ITEM mdk="672856" name="Public Sector Economics & Finance" relevant="376.0" />
</CATEGORY>
- <KEYWORD count="291">
  <KEYWORD_LIST>accountability, Accounting, accounting framework, accounting procedures, acquisition, acquisitions,
  ADB, adverse selection, affiliate, affiliated companies, Agriculture, applications, asset recovery, auction, auctions,
  Auditing, auditors, audits, authorized capital, autonomy, bank accounts, bank assets, bank capital, bank
  capitalization, bank lending, bank licensing, bank loans, Bank Supervision, bank supervisors, Banking Law, banking
  sector, banking services, banking stability, Banking Supervision, banking system, bankruptcy, bankruptcy cases,
  banks, bid, bidders, board members, bonds, borrowing, business enterprises, business entities, Capital Adequacy,
  capital markets, capital requirements, capital standards, CD, central bank, Clearing House, collateral, commercial
  banks, Company, compulsory insurance, confidential information, conflict of interest, consolidated supervision,
  consolidation, contingent liabilities, contractual relationships, Corporate bonds, corporate governance, corporate
  governance principles, corporate securities, corporate transactions, correspondent accounts, court, court fees,
  coverage, CPSS, credit institutions, credit risk, credit risks, credit unions, Debt, debt relief, decision making, deposit
```

Automatically Generated Metadata

foreign investment, fraud, gambling, GDP, governance arrangements, government bonds, government guarantees, government securities, growth rate, guidelines, income, incorporated, individuals, inefficiency, inflation, inflation rate, insolvency, insolvent banks, inspections, institutional impediments, institutional infrastructure, institutional structure, Insurance, Insurance companies, insurance system, insurances, integrity, interbank lending, interbank payments, interest rate, interest rates, interested parties, internal controls, International Accounting Standards, international banking, international capital, international capital markets, International Development, International Finance, International Finance Corporation, International Financial Institutions, International Monetary Fund, International Trade, investigation, investigations, Joint Stock Companies, laws, legal entities, legal entity, legal framework, legal protection, legal provisions, legislation, legislative framework, Liability, Liability Law, licenses, life insurance, limited, liquid assets, liquidation, liquidity, liquidity management, loan loss provisions, macroeconomic performance, macroeconomic stability, macroeconomic stabilization, management information systems, managers, market risk, market trading, maturities, Microfinance, monetary aggregates, monetary conditions, monetary policy, Money Laundering, money supply, moral hazard, National Bank, new entrants, nominal interest rates, nonbank financial institutions, Nonperforming loans, offerings, Open Market Operations, operational risk, organizational structure, own funds, ownership shares, partnership, Payment Systems, penalties, Pension Funds, Pensions, portfolio, portfolios, price stability, private capital, private ownership, privatization, profit margin, profitability, programs, provisioning, prudential rules, Public, public borrowing, public comments, public debt, public debt management, public infrastructure, Public Investment, Public Investment Program, public servants, Real GDP, real interest rates, recapitalization, recession, reciprocity, regulatory agencies, regulatory authorities, regulatory authority, regulatory framework, regulatory policy, reorganization, reserve requirement, reserve requirements, reserves, Return on Assets, revenue performance, risk management, risk-weighted assets, Savings, savings deposits, Securities, Securities Markets, securities transactions, shareholders, shareholding, small banks, social security, solvency, stakeholders, Statement, Stock Exchange, stocks, strategic investors, stress testing, subsidiaries, supervisory authorities, supervisory framework, sustainability, swaps, systemic risk, Technical Assistance, Tier 1 capital, trading, transition economies, transition economy, Transparency, treasury, treasury bills, Union, valuation of fixed assets, volatility, vulnerability, WTO</KEYWORD_LIST>

</KEYWORD>

</RESPONSE>

</TERAGRAM_XML>

Organizations Referenced

World Bank Org Names Edited Profile.tk2 - Teragram TK240

File Edit View Build Project Category Concept Testing Server Help

Teragram Org Names Edited Profile

- English
 - Concepts
 - Top
 - Governmental Organizations
 - IGOs**
 - NGOs
 - Other Organization
 - Public Companies
 - Universities

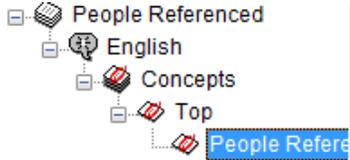
Bank for International Settlements,
BCEAO,
BIPM,
CAB International,
Caribbean Community,
Caribbean Community and Common Market,
Caribbean Export Development Agency,
CARICOM,
CCAMLR,
CEDAW,
Central American Bank for Economic Integration,
Central American Parliament,
Central Asian Cooperation Organization,
Central Bank of West African States,
Centre on Integrated Rural Development for Asia and the Pacific,
CERN,
CGIAR Publications,
Chemical Weapons Convention,
OPCW,
CIS,
Commission for Environmental Cooperation,
Commission for Labor Cooperation,
Commission for the Conservation of Antarctic Marine Living Resources,
Common Market for Eastern and Southern Africa,
Commonwealth of Independent States,
Commonwealth of Nations,
Commonwealth Trade Union Council,
Community of Portuguese Language Countries,

Syntax Check Classifier Load Text... Ln 14
 Grammar
 Filename

Taxonomy Dependencies Definition Testing Data Document

Ready NUM

People Profile With Authority File of First Names



```
# ROOT=*Name
# This profile is modeled on the new Person Name profile, bas

*Name = *FN1 #cap
*Name = *FN1 #cap #cap
*Name = *FN1 #cap - #cap
*Name = *FN1 *FN1 #cap
*Name = *FN1 _MIDDLEINITIAL #cap
*Name = _MIDDLEINITIAL _MIDDLEINITIAL #cap
*Name = _MIDDLEINITIAL _MIDDLEINITIAL _MIDDLEINITIAL #cap
*Name = *FN1 De #cap
*Name = *FN1 de #cap
*Name = *FN1 da #cap
*Name = *FN1 Da #cap
*Name = *FN1 de la #cap
*Name = *FN1 De la #cap
*Name = *FN1 Del Mar #cap
*Name = *FN1 du #cap
*Name = *FN1 du #cap
*Name = *FN1 du #cap
*Name = *FN1 von #cap
*Name = *FN1 ibn #cap
*Name = *FN1 ben #cap
*Name = *FN1 von #cap
*Name = *FN1 de #cap
*Name = *FN1 van #cap
*Name = *FN1 van de #cap
*Name = *FN1 van der #cap
*Name = *FN1 al #cap
*Name = Mr. #cap
*Name = Mrs. #cap
*Name = Ms. #cap
*Name = Miss #cap
*Name = M. #cap
*Name = Mme. #cap
*Name = Me. #cap
*Name = Mr #cap
*Name = Mrs #cap
*Name = Ms #cap
*Name = Mme #cap
*Name = Me #cap
```

```
# Be certain to include name
*FN1 = *FN
*FN = Ā, 'Kabaila
*FN = Aadam
*FN = Aadarshini
*FN = Aadeel
*FN = Aadi
*FN = Aadil
*FN = Aadilah
*FN = Aaditya
*FN = AE' amonn
*FN = Aafke
*FN = Aafreeda
*FN = Aage
*FN = Aaghaa
*FN = Aakanksha
*FN = Aakarshan
*FN = Aakif
*FN = Aalam
*FN = Aaleyah
*FN = Aalif
*FN = Aalim
*FN = Aaliyah
*FN = Aamaal
*FN = Aamani
*FN = Aamil
*FN = Aamina
*FN = Aamir
*FN = Aanchal
*FN = Aaqaa
*FN = Aaraa
*FN = Aaralyn
*FN = Aarif
*FN = Aariz
*FN = Aaron
*FN = Aarre
*FN = Aart
*FN = Aarthy
*FN = Aarti
*FN = Aaryn
*FN = Aasaf
*FN = Aashish
*FN = Aashiyana
*FN = Aashka
```



Grammar concept extraction allows us to define concepts based on semantic language patterns.



- [-] MARC Edition Statement
 - [-] English
 - [-] Concepts
 - [-] Top
 - [-] Edition

Fourteenth revised level,14th rev. level
 Fifteenth revised level,15th rev. level
 First revised update,1st rev. update
 Second revised update,2nd rev. update
 Third revised update,3rd rev. update
 3rd revised update,3rd rev. update
 Third revised update,3rd rev. update
 Fourth revised update,4th rev. update
 Fifth revised update,5th rev. update
 Sixth revised update,6th rev. update
 Seventh revised update,7th rev. update
 Eighth revised update,8th rev. update
 Ninth revised update,9th rev. update
 Tenth revised update,10th rev. update
 Eleventh revised update,11th rev. update
 Twelfth revised update,12 rev. update
 Thirteenth revised update,13th rev. update
 Fourteenth revised update,14th rev. update
 Fifteenth revised update,15th rev. update
 New Edition,New ed.
 New Ed.,New ed.
 New Issue,New issue
 New Release,New release
 New Level,New level
 New Update,New update

Syntax Check

 Classifier

 Grammar

 Filename

Load Text...

 Taxonomy Dependencies

 Definition Testing Data Document

ISBN Concept Extraction Profile

The screenshot displays the Teragram TK240 software interface for configuring an ISBN concept extraction profile. The window title is "ISBN.tk2 - Teragram TK240". The menu bar includes File, Edit, View, Build, Project, Category, Concept, Testing, Server, and Help. The toolbar contains icons for file operations and help.

The left pane shows a hierarchical tree structure:

- ISBN
 - English
 - Concepts
 - Top
 - ISBN

The main pane displays the REGEX configuration for the selected concept:

```
REGEX
ISBN. [7890] {3} [0-9] {9}, ISBN
ISBN. [0-9] {9}, ISBN
ISBN. [7890] {3} [0-9] {8} [xX] {0,1}, ISBN
ISBN. [0-9] {9} [xX] {0,1}, ISBN
ISBN. [7890] {3} [0-9] {9} [xX] {0,1}, ISBN
ISBN. [7890] {3} [-0-9 ] {10,14} [xX] {0,1}, ISBN
ISBN. [-0-9 ] {10,14} [xX] {0,1}, ISBN
```

The bottom pane contains a "Syntax Check" button and radio button options for Classifier (selected), Grammar, and Filename. A "Load Text..." button is also present. The bottom status bar shows "Ready" and "NUM".

Sentiment Analysis Profile

Sentiment Profile.tk2 - Teragram TK240

File Edit View Build Project Category Concept Testing Server Help

Sentiment Profile

- English
 - Categorizer
 - Top
 - Negative Sentiment
 - Positive Sentiment

OR

- DIST_200
 - "World Bank"
 - "abandoned"
- DIST_200
 - "World Bank"
 - "abashed"
- DIST_200
 - "World Bank"
 - "aberrant"
- DIST_200
 - "World Bank"
 - "abhorrent"
- DIST_200
 - "World Bank"
 - "abject"
- DIST_200
 - "World Bank"
 - "abjure"
- DIST_200
 - "World Bank"
 - "abolish"
- DIST_200
 - "World Bank"
 - "abolishing"
- DIST_200
 - "World Bank"
 - "abortive"
- DIST_200
 - "World Bank"
 - "absurd"

Sample Sentiment Output

```
<?xml version="1.0" encoding="utf-8" standalone="yes" ?>
- <TERAGRAM_XML>
- <RESPONSE>
  - <ENVIRONMENT>
    <SOURCE name="P:\Data Management\Teragram\Teragram Files from M Drive\TeragramIRIS\Secondary Education\Policy0Note1April2002.pdf"
      type="file" />
    <CATEGORY_SET name="Sentiment">Sentiment</CATEGORY_SET>
  </ENVIRONMENT>
  <LANGUAGE codepage="ASCII" name="ENGLISH" />
- <CATEGORY name="Sentiment">
  <CATEGORY_ITEM mdk="na" name="Positive Sentiment" relevant="281.0" />
  <CATEGORY_ITEM mdk="na" name="Negative Sentiment" relevant="92.0" />
</CATEGORY>
  <ORIGINAL_TEXT_EXTRACT>Information and Communication Technologies in General and Initial Vocational Education Towards A Coherent National Strategy
  Development March, 2002 Human Development Sector Unit Europe and Central Asia Region Document of the World Bank
  Introduc</ORIGINAL_TEXT_EXTRACT>
</RESPONSE>
</TERAGRAM_XML>
```

Country Categorization and City Extraction

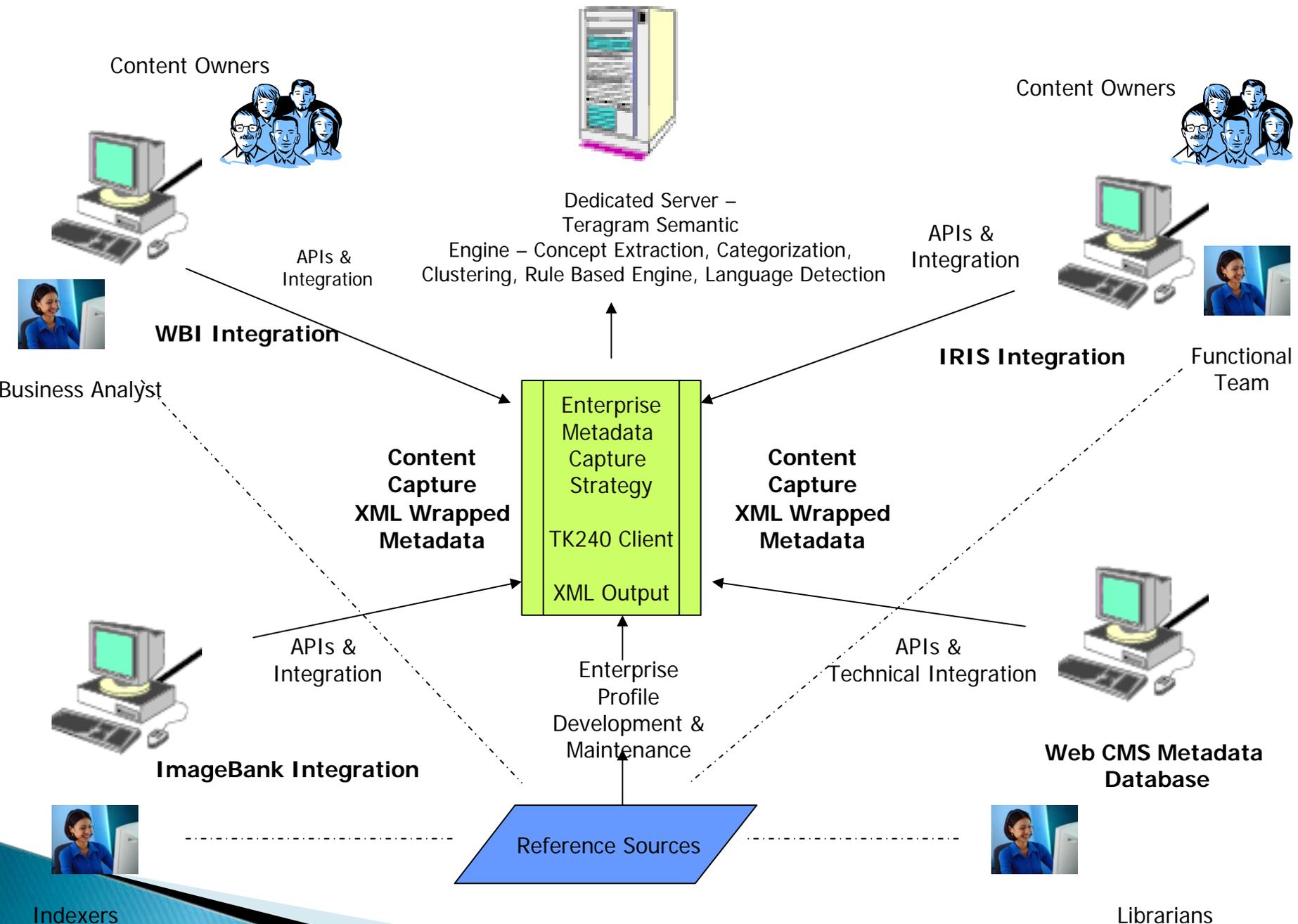
The image shows a software application window titled "Extended Country Profile.tk2 - Teragram TK240". The interface includes a menu bar (File, Edit, View, Build, Project, Category, Concept, Testing) and a toolbar with various icons. The main area is divided into two panes.

The left pane displays a hierarchical tree structure under "Extended Country Profile":

- English
 - Categorizer
 - Top
 - Country
 - 121825 - Stateless
 - 319022 - Anguilla
 - 439302 - Bouvet Island
 - 439305 - Western Sahara
 - 439306 - Faroe Islands
 - 439307 - Greenland
 - 439308 - S. Georgia and S. Sandwich Is
 - 439309 - Heard and McDonald Islands
 - 439310 - British Indian Ocean Territory
 - 439311 - Norfolk Island
 - 439312 - Nauru
 - 439313 - Niue Islands
 - 439314 - St. Pierre & Miquelon
 - 439315 - Pitcairn Island
 - 439316 - Svalbard and Jan Mayen Islan
 - 439317 - San Marino
 - 439318 - French Southern Territories
 - 439319 - Antarctica
 - 439320 - Tokelau
 - 439321 - USA Minor Outlying Islands
 - 439322 - Holy See (Vatican City State)
 - 439323 - Virgin Islands (UK)
 - 439324 - Wallis and Futuna Islands
 - 439325 - Mayotte
 - 474038 - Kosovo
 - 82500 - New Zealand
 - 82501 - Oman
 - 82502 - Panama
 - 82503 - Peru
 - 82504 - French Polynesia
 - 82505 - Papua New Guinea
 - 82506 - Philippines
 - 82507 - Pakistan

The right pane displays a list of city names, organized into two groups under "AND":

- OR
 - "Brazil"
 - "Brasil"
 - "Republica Federativa do Brasil"
 - "Federative Republic of Brazil"
 - "Republique federative du Brasil"
 - "Brazilia"
 - "Brasile"
- OR
 - "Abacaxis"
 - "Abadia dos Dourados"
 - "Abadiânia"
 - "Abaeté"
 - "Abarracamento"
 - "Abatiá"
 - "Abel"
 - "Abre Campo"
 - "Abreu e Lima"
 - "Abufari"
 - "Abunã"
 - "Acajutiba"
 - "Acará"
 - "Acarã"
 - "Acará-Mirim"
 - "Acarauá"
 - "Acarauá"
 - "Acari"
 - "Acorizal"
 - "Acre"
 - "Acre"
 - "Açu"
 - "Açuã"
 - "Açucena"
 - "Açude Aracatiçu"
 - "Açude Araras"
 - "Açude Banabulú"
 - "Açude Orós"
 - "Açude Poçoda Cruz"
 - "Acurauá"
 - "Adamantina"
 - "Adrianópolis"



Teragram Architecture – 50,000 Foot View