This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

Lessons Learned in Content Architecture Harmonization and Metadata Models

Shana Wagger, Consultant, The World Bank, and Senior Editor, Alexander Street Press; swagger@gmail.com; 1-202-210-0774; 3025 Ontario Rd, NW, #302, WDC, 20009, USA
Randi Park, Publishing Officer, The World Bank; rpark@worldbank.org; 1-301-704-1665; 1 Baederwood Court, Derwood MD, 20855, USA
Denise Ann Dowding Bedford, Ph.D., Senior Information Officer, The World Bank, and Adjunct Faculty, IAKM Program, Kent State University; dbedford@worldbank.org; 1-202-458-1927; The World Bank, MSN I4-400, 1818 H St, NW, WDC, 20433, USA

Keywords: Classification Schemes, Content Architecture, Metadata Strategies, Product Design

I. Overview of *The World Development Report* Series and Project

The World Development Report, published by the World Bank first in 1978 and annually since then, is one of the most influential references on the world economy and the state of economic and social development. Each year, a new team of World Bank authors tackles a specific development topic, from agriculture, poverty, health, and infrastructure to spatial geography and development issues for youth and the next generation. This flagship finds its way into many hands and institutions, having a wider and broader readership than many of The World Bank's other publications.

For the past few years, the *Reports* had been available in a number of print and electronic formats, including a rudimentary Omnibus CD holding multiple editions, but there had never been a concerted effort to design a special platform for the series, one that could evolve and migrate to the web.

Using the occasion of the thirtieth anniversary of the *Report* as an opportunity to take the series a step further, The Office of the Publisher, with the help and advice of the Development Data Research Group, began planning a new electronic product to make the entire series discoverable and searchable within a single interface and platform. This entailed assessing the content base as a whole, converting material to XML and other formats, designing specific search and browse strategies, developing the architecture and interface, and creating and working with metadata. These initiatives, many of which were new to the Bank, are discussed below, along with some recommendations and strategies for future releases.

The Complete World Development Report: 1978 to 2009 in DVD form came to fruition in late 2008. It is slated to become an online product in 2009.

Figure 1

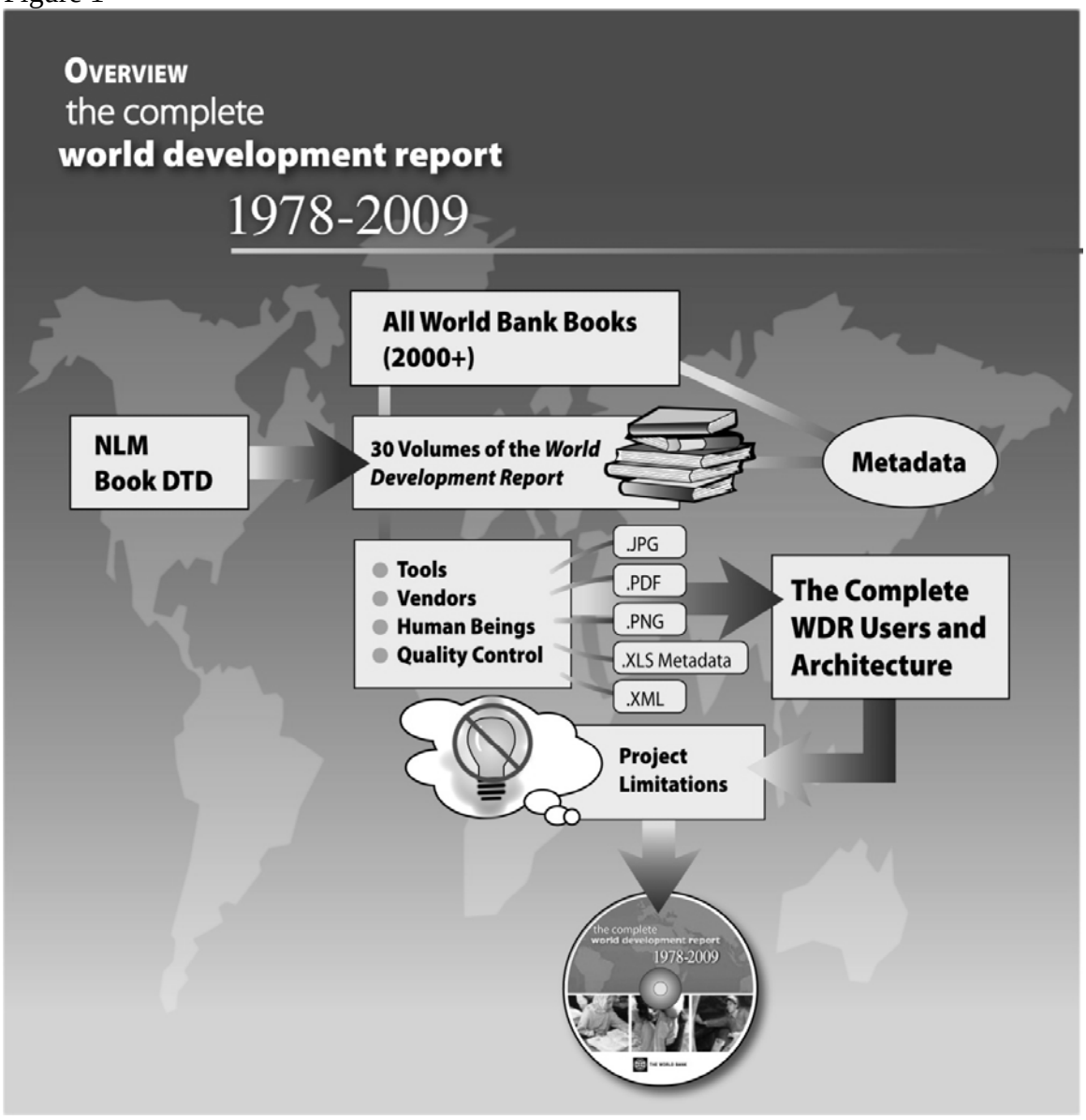

Figure 1: Overview of the workflow, process, and tools

II. Identifying current and future goals and needs

The Complete World Development Report project focused on that content pool, of course, but the project required the team to take stock of other content as well as the overall goals, needs, and business and publishing plans:

- The goal of raising awareness of the problem of poverty, the need for social and economic development, and what Bank knowledge, programs, and involvement are available are evergreen initiatives, both for the Bank and for the series itself.
- In terms of the *World Development Report* series itself, first and foremost were rejuvenating the content to bring its insights and historical views to a wider

audience and preserving the content in a viable and neutral format, ready for reuse.

Other key goals focused on the following:

- Conversion, atomization, and aggregation had to be viable and replicable in terms of the publishing program at large, rather than provide a solution for a single project and content type.
- Any converted content had to be reusable for future projects and in conjunction with other legacy and yet-to-be published content.
- Architecture, functions, and tools had to create a consistent and evolving approach for our users and our overarching publishing program, including any migration from static to dynamic electronic platforms.
- Metadata models and semantics had to take into account future uses and integration of content.
- Any scheme for ids, DOIs, and reference tagging had to address the needs of current and future partners, like crossref.org.
- Any new base product and platform had to take other truths into account: As content becomes accessible through one or more portals, its reach is broader and its audience less specialized. In this environment, users have become accustomed to more control as well. This meant providing data mining tools allowing users to customize results, rather than simply providing the results themselves.

III. Analyzing the series

Determining formats and availability of legacy content

Published beginning in 1978, the thirty-book series had run long enough to span several distinct eras in publishing and storage technologies, a few different philosophies on saving the backlist, and an evolving relationship with a copublisher. The earliest years were available only in low-quality black-and-white photocopies and image files, neither of which could be scanned easily for any kind of quality conversion. The latest years were available in easily convertible application files and high-resolution PDFs. The middle years fell in between these two extremes.

The project required three plans for conversion of the series to XML and PDF and two different vendors, based on the range of formats and the different scanning or keyboarding arrangements required by the variation in quality.

Understanding the evolution of series content

Several dimensions were critical to decisions about content architecture and product design, and these could only be assessed on a book by book basis. Not only had the series evolved over thirty years, but each book, by design and artifact, was unique. The authoring team, topic, and focus changed for each *Report*, so that books reflected little consistency in structure and elements, but rather author preference and understanding.

The need to provide background on the World Bank itself, the global economic situation, and historical issues or events before delving into the chosen topic also evolved, requiring less coverage and emphasis as the series grew more recent. This kind of background material posed a special challenge in our metadata model.

The mix of historical coverage, topical coverage, and data changed throughout the Series. As historical coverage lessened, inclusion of data increased, but the data themselves were not normalized or comparable from year to year until quite late in the series.

Early *Reports* predated the ISBN, and most reports predated the DOI as well. This was important because these numbers would be key to the construction of all unique identifiers throughout the XML files.

Assessing features and structure

Part and chapter structure varied greatly over the series, with different ways of using and incorporating Introductions and Overviews, including at least one Overview that actually concluded, rather than began, a book.

A handful of recent books used specially designed, highly focused short features that stood alone, outside the chapter structure and concept. Each set of features had a unique role and nomenclature, but together created a class of element that needed special conversion instructions and incorporation into the architecture and search strategy.

Cross-reference styles varied greatly from book to book, including references by page number, "previous" and "next" section callouts, and chapter number. Only the last kind of cross reference could be programmatically linked in the conversion process, and because these links would be a critical way for users to follow train of thought and access content, we knew that our own post-conversion work needed to ensure these other cross references were made dynamic.

Citations, endnotes, and references changed dramatically through the life of the series, in terms of elements, style, organization, and placement, posing untold challenges for the conversion houses and for us in the consistent rendering, interlinking, and architecture for these elements. Ensuring tagging that allowed us to work with partners like crossref.org was important, but also a challenge with this content pool.

Another very unusual challenge was the nomenclature used in certain special appendices, like those in the 2009 *Report*. Here, the exact table title was reused as a heading encompassing the source list for that table. While context and distinction between these identical titles was clear in a print environment, they were very confusing, and looked like errors, in the dynamic Table of Contents. It became even more problematic when the less relevant heading was included as a link in the dynamic Table of Contents, but the table itself was not. (See below.)

Documentation for the conversion houses included baseline instructions for the series as a whole, a book-by-book analysis of anomalies, and a spreadsheet setting out how ISBNs and DOIs should be used and interpreted for ids and e-location ids through the XML and image files.

An early solution leads to late problems

At least two copies of each *Report* were necessary for the project. Because of the complexity of formats and solutions, and the need to move quickly, once two complete sets of books had been identified, one set was shipped off to the conversion houses and the other kept at the Bank for analysis and reference.

It came to light during the post-conversion quality control process that a number of *Reports* had actually existed in two versions, one older version bearing a Bank ISBN and a later copublished version bearing the copublisher's ISBN. Unknowingly, we had really sent one edition for conversion and kept the other at the Bank. Because our DTD/XML convention was to use the DOIs and ISBNs as the basis for the thousands of e-location and other ids throughout each converted file and its images, the files sometimes reflected the wrong set of numbers as the base id. (See later discussion of this issue.)

This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

IV. Decisions about the content pool

Including the entire series of the *World Development Report* in the project was a given, but was not construed as a limitation on thinking about the content.

Should all the content in each *Report* be included and included in the same way? Was there any other content to be created or integrated? Was there other content to be wrapped into a later edition, and if so, what needed to be taken into account with this early version? These were some of the questions to be considered and either acted upon or noted for future development.

When it came to the series content, it was clear that the historical and substantive analysis was of high value, and the conclusion was that all chapters and features should be converted and integrated in some way into the strategies for discovery. In addition, abstracts were selected for inclusion so users had a richer description of content than a simple taxonomy or set of titles could provide.

When it came to the data appendixes that started to flow into the series midway, it was an important realization that the earlier data were not yet comparable or normalized, possibly misstating comparisons that would be made from year to year. In addition, users had taken many steps forward in terms of working with data and expecting a certain standard of data from the World Bank. There were a few key conclusions here:

- A dynamic statistical database with normalized and comparable *World Development Indicators* data matching the frame of reference for the Series would better serve users interested in the data. This became an important global navigation feature in the product.
- To keep confusion to a minimum, the database would replace the data appendixes, which would not be converted or integrated into the architecture.
- For those users interested in the *Reports* and their data as artifacts, a PDF of each *Report* would be accessible on the landing page for that *Report.*

When it came to other back matter--not the data appendixes, but the epilogues, other appendix tables, bibliographies, endnotes, and reference lists—all content was slated for conversion. Only the very last few *Reports* included indexes, and we ended up with the worst of both worlds. Those few indexes were converted, but without dynamic links because we did not have the capacity to translate page numbers into locations in the XML, and without an indexing module for our DTD.

When it came to front matter, there was a range of content over the years, with dissimilar titles for very similar and standard features. Where the XML structure allowed us to generate something better dynamically—Tables of Content, for example--that original content was not converted. Anything descriptive or analytical was slated for conversion.

Figures, tables, boxes, and to some extent equations played an important role in the series. Any method of transformation or capture had to make the content discoverable within our metadata model as well as searchable dynamically by the user.

- For figures, all substantive information--labels, titles, captions, notes, and sources—was converted to XML, with the graphic itself a .jpg linked into the XML via unique id. This also allowed us to reuse figures and store them separately.
- Tables posed a spectrum of choices, including the simplest treatment as a figure, as described above. The more complex conversion to XML was chosen, however, to allow us to build an “export data” feature in a future release. The conversion was a little painful, because of the size of the tables, the lack of standardization in terms of alignment, and the familiarity required to interpret some of the less matrix-based tables, which our vendors did not have.

- Treatment of equations was not necessarily straightforward either. On the one hand, our DTD included everything we needed: both display and inline formula elements and a MathML module for use with inline and display equations. On the other hand, only a subset of browsers could render MathML reliably, which could be a problem for our user base. The decision was to provide options down the line, and all equations were captured as both presentation MathML and a graphic .png file linked into the XML via unique id. (Because equations pose no issues with color or resolution, .png made more sense than .jpg.)

The timeframe of the series also meant that historical context, for both the Bank's activities and the world at large, particularly given the Bank's own genesis as a Bretton Woods institution, would be a valuable addition for users. Work began on revising and updating a World Bank timeline specifically for inclusion in the project.

The World Bank also has a somewhat specialized view of the world: It defines its regions in a certain way, and it focuses on developing countries much more than high-income and other donor countries. For example, a World Bank region like East Asia and the Pacific does not include Japan, which is not always apparent or easily understood by those outside the Bank, particularly as the audience becomes broader and less specialized. To help users understand the Bank concept of region, income, and country, additional information about these needed to be developed and integrated into the architecture.

Each *Report* is based on a rich series of background papers and notes, which were only available in a manipulable form for the very late years. Given the timeframe and budget, it was not possible to assess, convert, and integrate this level of content into the project. To prepare for this step, a new value of "background-paper" was built into our DTD and used to identify all core papers and notes cited throughout the series during the conversion process.

V. Defining strategies for discoverability

Four overarching concerns shaped our thinking:

- What was commonly expected by users researching any kind of World Bank content;
- What was specifically appropriate to the user for this content pool;
- What tools and functions could be modified for future products and content pools;
- What we needed to learn, perhaps incrementally, about our metadata model.

And, of course, in the background were our budget and our timeframe --smaller, shorter, and more constrained for this first effort.

The intersection of the first two issues was the most complex because the *World Development Report* is actually one of the best known, but least typical Bank publications. For example, "region" is an important organizational principle and program focus at the Bank itself and a very common frame of reference for those researching Bank content. It should be a standard strategy for browsing and searching World Bank collections. Yet the *Reports* themselves were so broad-based that a regional browse in this context would have yielded the same list of reports (and chapters, most likely) for each region, a frustrating experience for users.

Another example centered on the inclusion and incorporation of the timeline. Again, on the surface, there seemed to be a clear tie between a year-by-year feature and an annual publication like the *Report*. Yet the year served as a publishing cycle determinant for the *Reports*, not as a content focus. An architecture that drew a close link between the *WDR 1993* and the timeline for 1993, for example, would have looked

appropriate superficially, but it would have drawn researchers to make misleading conclusions about content and timeframe.

Impact of the titling convention

The series titling convention of *World Development Report [year]: [Topic or concept]* shaped part of the architecture.

On the one hand, it increased power because almost any function based on the title served double duty for chronology, because the year was the first distinguishing term in each report title.

On the other hand, it decreased power because it gave users less distinguishing information on which to search or to make decisions about substance and relevance.

Browsing Strategies

Because our metadata model had not yet been applied to or refined for content at the chapter level, we determined to keep the browse strategies simple for this experimental iteration.

Each browse in this project would take users initially to the full *Report*, rather than a chapter or subchapter passage. Users would be able to browse any portion of the *Report* they desired, but would always be presented with the landing page for that full *Report* rather than drilled into a particular chapter or subsection. This decision might not be repeated for other products or later versions of this product, once our metadata model is refined.

Browse by Title. This standard and expected browse allowed users to see the entire list of series titles in reverse chronological order. It was more of a chronological browse, but we labeled it "Browse by Title" because that convention was clear as well.

Browse by Topic: This standard and expected browse of 20 to 25 major topics expanded to show the list of relevant *Reports* under each topic. Each *Report* was assigned no more than two major topics, and each list of *Reports* within a topic could be resorted from newest to oldest.

Browse by Chronology: This standard and expected browse was implemented as part of the Browse by Title above, rather than as an individual feature. This was a decision unique to this content pool and unlikely to be repeated in another product.

Browse by Region: This standard and expected browse was not built for this content pool, as noted above, because the *Reports* were equally associated with each region. To help answer questions about the regions, and to preserve a place in the interface for such a browse when the content pool benefited from it, the global navigation included a screen for Regions, providing maps and information about countries in each region and income levels. Future versions of the project could include a regional browse, but such a browse could only be invoked at a chapter or section level and would not be applicable to all chapters or sections.

Browse by Country: This standard and expected browse was not built for this content pool because much deeper analysis of content was required to understand the relevant associations. Future versions of the project might include a country-based browse, but such a browse would apply to much less content as well as much more granular content. In addition, user expectation of a nested relationship from region to country would need to be addressed.

Searching Strategies

Although our metadata model had not yet been tested for chapter-level associations, as noted above, it was seen as less of an impediment to searching at the chapter level, as long as we could build in some good data-mining tools for users to apply

to their results. To keep search results focused, searches would return only certain categories of HTML content:

- Any chapter in a *Report*, whether numbered or unnumbered
- Any part introductory text
- Any special feature that fell within the body, but outside the chapter structure, of a *Report*
- Any Introduction or Overview

Excluded content included any front matter; and any back matter, including reference lists, endnotes, appendix tables, and other content. All excluded content remained browsable, however, and searchable through "cntrl f" functionality whenever the user was viewing *Report* content. The decision to exclude back matter was particularly important, because a full text search for almost any term, country, region, or topic would most likely return all 30 sets of back matter simply because of the lengthy, 1000+ item reference lists concluding most *Reports*.

PDFs were also excluded from searching or browsing in the interface, but could be browsed and searched through Adobe.

One thing we could not build this time, but would like to build for future products, is the ability for users to choose what content type they'd like to search. Because our DTD was selected and customized with this in mind, we should be able to build, for example, a search that allows users to find an editor or an author of a specific type of citation within only reference lists, or to find a particular word or concept within figures, and so forth. This would be powerful, but could not be implemented at this time.

The fields for the advanced search included these parameters, again shaped to some extent by the titling convention of the series:

- Free text box for searching any words or phrases in any of the searchable HTML content. This included figure captions and metadata, but not words or phrases used in the graphic itself.
- Checkbox list of titles, in reverse chronological order. This was a unique construction, mandated by the titling convention of the series, which left too few words for a useful implementation of the more standard free text box for title. Of course, it served double duty for limiting to a timeframe. The default was all *Reports*. The unique structure of this search also had an impact on the Boolean relationships between advanced search fields, as noted below. (Future versions could include a free text box search for chapter titles, but this was not implemented for this initial version.)
- Checkbox list of topics: This was standard and expected. The default was all topics.
- Checkbox list of regions: This was standard and expected, and it worked to some degree at the chapter level, where it would not have worked at the *Report* level.
- Typical search fields like Author/Editor and Publisher were not included because they had no relevance in this content pool, with "World Bank" as the only publisher and author provided. This was unique to this content pool.

Boolean logic: The most powerful search logic would have been to AND all the user's choices, so that, for example, the user could find all chapters published since 1995 about both Health and Gender issues. The construction of the Title search option, however, made AND an impossibility because a chapter could be associated with multiple regions, or multiple topics, but only one *Report*. In this regard, the default logic became OR within a search field (i.e., it must match at least one of the multiple topics selected) but AND between search fields (i.e., it must match at least one topic selected AND at least one region selected).

The search results were constructed to give users multiple ways to limit and sort their own search results:

- Default presentation was by relevance.
- Users could resort by *Report* title (also chronology).
- Users could limit results by any topic or region used in their search. (In future products, we would like to build this out to provide limiting by any topic or region associated with any returned content, no matter what the users selected, as well as limiting or sorting by type of content [e.g., chapters, special features, individual figures, and so forth].)

Clicking on any result took users to the document page showing that chapter. Special navigation at the top of that screen allowed users to see previous and next search results on that refreshed document screen rather than having to toggle between the results screen and the results themselves.

Content presentation and linking

Each *Report* had a landing page that linked into a document screen.

The landing page, as described above, was accessible via any browse and included the abstract, a linked Table of Contents, bibliographic details, and the PDF. (Author and publisher details were included as bibliographic details, even though it meant repeating "World Bank" in every instance for both.) Clicking any part of the Table of Contents refreshed the window with the document screen.

The document screen refreshed to show the selected chapter of the *Report*. It also included a hyperlinked Table of Contents showing every part, chapter, and level of heading and subheading. This listing, however, did not include table, figure, or box titles in situ. These were relegated to a separate hyperlinked listing, although the tables, figures, and boxes themselves were readable in context of their chapter.

A special feature let users toggle endnotes off and on in situ in the content, a workaround for having chapter-by-chapter endnotes collected at the end of each *Report*, rather than as back matter for each chapter.

There was no elegant solution for author/date citations in the text or in the endnotes, all of which jumped the user to the back matter, where all citations resided.

These are all things the team would like to address in a new release.

VI. Metadata Requirements

Metadata were an essential component of the *World Development Report* information architecture. Metadata were required to support the browse and search functions. As users could access information at several different levels of aggregation, metadata had to be available for "whole" objects and their "parts" (i.e., metadata had to be available at the report level, the chapter level, for graphics and for tables).

Two types of metadata were required to support the information architecture – at each level of aggregation:

- Basic descriptive information (author, title, date, type of object)
- Classification information (i.e., topic, country, region classification)

Information professionals have a wealth of bibliographic principles and rules, classification schemes, authority control sources, controlled vocabularies and thesauri to draw from in generating metadata. For the *WDR* project, the team looked to the information science tradition of analytic description. Analytic cataloging produces metadata for parts, typically journal articles or volumes in a series, with an overarching

metadata record for the series or serial title. The project team wanted to understand how well standard practices worked at both the whole and the part level for both types of metadata.

General Descriptive Metadata

Generally, we found that descriptive cataloging rules (author, title, object/resource and date) worked equally well at the report, the chapter or the object level. Basic descriptive cataloging rules worked well for each level of information aggregation. Rules for identifying authors, titles, and dates held regardless of whether they were applied to the report, the chapter, the table or the graphic. This aligned with expectations.

Topic Classification of *WDR* DVD Product

The project team's first classification challenge was selecting a topic classification scheme that was well suited to the product and the audience. Three topic classification schemes were available:

- Economic Sectors and Thematic Areas
- World Bank Enterprise Topic Classification Scheme
- World Bank Office of the Publisher's subject listing

The selected topic scheme had to meet several criteria:

- Provide good coverage of thirty years of changing economic development and policy topics;
- Be easily understood by the general public, subject matter experts in any of the potential thirty high level knowledge domains of interest to the Bank over those thirty years;
- Bridge the thinking of economists over thirty years of economic development learning;
- Provide granularity of topics that would work equally well at the whole and the part level.

The first option was the Economic Sectors and Thematic areas. This was not recommended for several reasons. First, sectors are defined as the economic sector of the economy which are impacted by the development work. The scheme is intended to provide a project finance and performance view. As a result, the baseline scheme is economics-oriented, and varies from good practices for forming or managing a classification scheme. It is not comprehensive of all topics of interest to the Bank over the past thirty years. Furthermore, the listing of sectors is not supported by consistent scope definitions and tracings. Guidance for consistent interpretation by either human or machine classifiers is not provided. Enhancing the scheme for use in the project would have been a significant effort of its own. This scheme was not an optimal choice.

The second option was the World Bank's Enterprise Topic Classification Scheme. This scheme is intended for use in large collections. It is inclusive of all Bank interests over the past sixty years. It is also intended to be sufficiently flexible and robust to manage those interests in the future. This scheme is comprised of two levels – Topics and Subtopics. There are 30 Topics, and approximately 750 Subtopics (Figure 2).

Figure 2. Enterprise Topic Classification Scheme – Top Level

- Agriculture
- Communities and Human Settlements
- Conflict and Development
- Culture and Development
- Education
- Energy
- Environment
- Finance and Financial Sector Development
- Gender
- Governance
- Health, Nutrition and Population
- Industry
- Information and Communication Technology
- Infrastructure Economics and Finance
- International Economics and Trade
- Law and Development
- Macroeconomics and Economic Growth
- Poverty Reduction
- Private Sector Development
- Public Sector Development
- Rural Development
- Science and Technology Development
- Social Development
- Transport
- Urban Development
- Water Resources
- Water Supply and Sanitation

Figure 2: Enterprise Topic Classification Scheme – Top Level

The schema is governed and managed according to a set of institutional Guiding Principles and Best Practices. The Guiding Principles and Best Practices govern the structure, its behavior and use in institutional applications, business rules, and change management procedures. The schema is applied to all the Bank's major information repositories to support browse and search. It is intended for use in a large collection of information (10 million or more documents). The scheme is used for all types of content, including large 500 page research papers, publications, short communications and email messages, project documents, and so forth. It provides a level of interoperability and consistency across the organization. The Enterprise Scheme has principles, business rules and an active governance model.

One of the Guiding Principles is that each level of the schema is held to 30 classes. There are 29 top-level Topics (Figure 2). Each top-level Topic may have up to 30 subtopics. Figure 3 illustrates the subtopics defined for Health, Nutrition and Population.

The scheme promotes stability and predictability at the top level, and flexibility and change at the second level. Content is always classified to the second level of the scheme, with the top level inferred from the second level. The second level may reflect different perspectives on a topic, areas of practice or implementation, or possibly a subdomain. Because of the nature of the Bank's work, and the changing needs of economic development, the second level classes are not entirely orthogonal from one another. The Bank has struggled with cross-cutting perspectives for several decades. This challenge has been met by supporting some degree of overlap in the definition and scope of subtopics. For example, the two subtopics – Climate Change – and – Biodiversity – may both have ecosystem cycles or mountain ecosystems as core components of the class definition. Classification of content to both subtopics would be supported because it would place the content in those places/classes where people browsing or searching would expect to find it.

The classification scheme is held to two levels. Efforts to refine the scheme to include a third level (sub-subtopics) have been opposed, as the third level naturally

become intensely cross-cutting. Rather than develop a further refined classification scheme, the Bank uses concept level indexing to complement the classification scheme.

Figure 3. Enterprise Topic Scheme Second Level for Health, Nutrition & Population

- Adolescent Health
- Alcohol and Substance Abuse
- Avian Flu
- Cholera
- Communicable Diseases
- Country Population Profiles
- Demographics
- Disease Control and Prevention
- Early Child and Children's Health
- Environment and Health
- Family Planning Research
- Food and Nutrition Policy
- Health and Poverty
- Health and Sanitation
- Health Economics and Finance
- Health Indicators
- Health Insurance
- Health Monitoring & Evaluation
- Health Policy and Management
- Health Project Design and Implementation
- Health Service Management and Delivery
- Health Systems Development and Reform
- HIV AIDS
- Immunizations
- Malaria
- Mental Health
- Nutrition
- Pharmaceuticals and Pharmacoeconomics
- Population and Development
- Population Policies
- Public Health Promotion
- Reproductive Health
- SARS
- Tobacco use and Control
- Tuberculosis

Figure 3: Enterprise Topic Scheme Second Level for Health, Nutrition & Population

Figure 4. The Office of the Publisher's Subject List

- Agriculture and Rural Development
- Banking Finance and Investment
- Business Procurement
- Commodities, Pricing and Trade
- Current Affairs
- Development Economics
- General Economics
- Education and Training
- Energy, Industry and Mining
- Environment, Pollution Prevention
- Gender
- Globalization
- Governance, Civil Society and Participation
- Health, Nutrition and Population
- Infrastructure, Transport, and Urban Development
- Labor and Income
- Legal and Judicial Issues
- Poverty
- Private Sector
- Public Policy
- Social and Cultural Issues
- Technology and Telecommunications
- Water Supply and Sanitation
- World Bank

Figure 4: The Office of the Publisher's Subject List

This scheme was not selected for use in the *WDR* DVD product because the project team's near term goal was to align the *WDR* DVD content with the Office of the Publisher's other publications and products. At the time the product was released, those products were classified to a third option, the Office of the Publisher's Subject List. The Subject List is a flat listing (single level) of twenty-four subjects (Figure 4). The subjects provide broad subject coverage, and align with the Enterprise Topic Scheme on several points.

The project team found that the single level subject list was a good fit for the browsing and searching needs based on the number of reports and chapters in the DVD product. As the *WDR* content is integrated into the larger information repositories at the World Bank, the content has been classified to the Enterprise Topic Scheme.

Topic Classification of *WDR* Products in a Larger Context

The project team was interested in understanding how *WDR* content would classify when included in the larger information context of the World Bank. At the present time, the World Development Reports are classified to the Enterprise Topic Scheme at the "whole" level. We set out to answer the question: would this approach work equally well for the "parts" or chapters? How would chapters classify when included in the larger collections and working with the enterprise-wide scheme? To answer this question, we classified the *WDR* "part" content using the same methods applied to all other World Bank information content.

Three strategies are used to generate metadata today: human-generated, machine-generated, and a hybrid of both human and machine assisted. Each approach has advantages and disadvantages for general use and specific to this project. Each strategy leverages the same components, but in different ways. A full description of the three methods is provided in Appendix A. The description is provided to explain how the project team examined the topic classification question.

The World Bank uses a machine-based method to classify information resources to the enterprise topic classification scheme. Information is classified at the subtopic level, with the topics inferred from the subtopic assignments. The World Bank adopted this approach five years ago because of the limitations observed in human classification. We found that:

- Professional classifiers who had a broad knowledge of the Bank's interests could do a very good job of identifying all of the subtopics addressed, but only at considerably higher costs and for a very small fraction of the information generated by the organization;
- Non-professional classifiers automatically selected topics that were of interest to them but did not necessarily pertain to the content of the information being classified;
- When properly engineered (i.e., trained to think like a human classifier), a semantically-grounded machine classifier could do as well as or better than a person – and process the full set of information produced. Appendix A provides more detailed discussion of the semantically-grounded technology, and how it is trained.

Evaluation and Observations

The World Development Reports and the individual chapters were classified to the Enterprise Topic Scheme at the subtopic level. The topics were inferred from the subtopics. While further testing is required before we can conclude that this practice would apply to "whole" and "parts" in all cases, we observed that:

- The full report classified with a high degree of relevancy or 'goodness of fit' to between five and seven top-level Topics (Figure 2).
- These top-level Topics broken down represented ten to twenty second-level Subtopics (Figure 3).

For example, the *World Development Report 2005: Better Investment Climate for Everyone* classified to the following subtopics and topics (Table 1):

Table 1.
WDR 2005 Classification to Topics and Subtopics (Full Report)

Environment	Finance and Financial Sector Development	Labor and Social Protections	Macroeconomics and Economic Growth	Trade and International Economics
Environmental Economics and Policies	Debt Markets	Labor Policies	Economic Theory and Research	Trade and Regional Integration
	Emerging Markets			
	Investment and Investment Climate			
	Access to Finance			
	Non-Bank Financial Institutions			
	Bankruptcy and Resolution of Financial Distress			

Table 1: WDR 2005 Classification to Topics and Subtopics (Full Report)

When classified individually, the *WDR 2005* chapters surfaced ten additional Topics and sixteen Subtopics. The additional classes were consistent with the coverage and provide important new access points for discovery in a larger collection of information. The additional subtopics and topics are described in Table 2:

This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

Table 2a.
WDR 2005 Classification to Topics and Subtopics (Chapter Level)

Communities and Human Habitats	Education	Gender	Information and Communication Technology	Infrastructure Economics
Common Property Resource Management	Tertiary Education	Gender and Law	e-Business	Private Participation in Infrastructure
			ICT Policy and Strategy	

Table 2b.
WDR 2005 Classification to Topics and Subtopics (Chapter Level)

Law and Development	Public Sector Development	Transport	Health, Nutrition and Population	Poverty Reduction
Legal Regulation and the Business Environment	Public Sector Economics and Finance	Transport Economics, Poliy and Planning	Health Monitoring and Evaluation	Pro-Poor Growth
Law and Gender	Public Sector Corruption and Anticorruption		Population Policies	
Trade Law				
Arbitration				

Table 2a & B: WDR 2005 Classification to Topics and Subtopics (Chapter Level)

A second example is the classification of the *World Development Report 2007: Development and the Next Generation*. The full report classified with a high degree of relevancy or 'goodness of fit' to six top-level Topics, and ten second-level Subtopics (Table 3):

Table 3a.
WDR 2007 Classification to Topics and Subtopics (Full Report)

Education	Finance and Financial Sector Development	Gender
Secondary Education	Access to Finance	Gender and Education
Primary Education		
Education for All		

Table 3b.
WDR 2007 Classification to Topics and Subtopics (Full Report)

Governance	Health, Nutrition and Population	Social Development
Youth and Governance	Population Policies	Street Children
	Adolescent Health	
	Health Monitoring and Evaluation	

Table 3a & b: WDR 2007 Classification to Topics and Subtopics (Full Report)

When classified individually, the *WDR 2007* chapters surfaced the same top-level Topics and an additional seven second-level Subtopics. The additional topics would be valuable for discovery in a larger collection of information. Specifically, additional Topics and Subtopics are listed in Tables 4a and 4b.

Table 4a.
WDR 2007 Classification to Topics and Subtopics (Chapter Level)

Communities and Human Habitats	Industry	Information and Communications Technology	Labor and Social Protections
Housing and Human Habitats	Technology Industry	e-Business	Child Labor

Table 4b.
WDR 2007 Classification to Topics and Subtopics (Chapter Level)

Poverty Reduction	Public Sector Development	Rural Development
Pro-Poor Growth	Public Sector Corruption and Anti-Corruption	Rural Poverty Reduction

Table 4a & b: WDR 2007 Classification to Topics and Subtopics (Chapter Level)

From these preliminary evaluations, it would seem that discovery in a larger collection is enhanced when:

- Chapters are individually classified and made accessible for direct discovery;
- Whole reports or whole objects infer or 'inherit up' the classes that are assigned at the chapter level.

If the classification of full reports or documents is inherited down to individual chapters two problems may result:

- Issues discussed in the parts will be overlooked;
- Users may be directed to individual chapters where coverage does not exist (the inheritance pattern does not hold).

The same test and evaluation was performed on indexing concepts (keywords as opposed to subject classes) at the chapter and the full report level. The impact of inheriting down vs. inferring up was more apparent at this finer level of indexing granularity. The complete set of chapter level indexing concepts would enhance user discoverability at the full report level. Taking only the report level concepts and inheriting them down to chapters, would generate the same type of errors noted in the topic classification.

We caution that these results are shared only as preliminary observations. Further work and testing needs to be completed before we can generally say that these practices would hold generally across content. However, the results are consistent with our general observations of access patterns.

In sum, in a larger context the classification of *WDR* reports and chapters can be supported by standard Bank classification procedures. The approach to defining classes for the full report, though, would likely change given what we have learned in this exercise. Given further testing, the project team would recommend that we classify parts and assign the sum of the part classes to the whole report. As a result of this exploratory work, the Bank's approach to classification and indexing of report series is changing from a top-down inheritance to a bottom-up inference.

VII. Appendix A: Three Approaches to Generating Metadata

Three strategies are used to generate metadata today: human-generated, machine-generated, and a hybrid of both human and machine assisted. Each approach has advantages and disadvantages for general use and specifically related to this project. Each strategy leverages the same methods and knowledge, but in different ways.

Human Generation

In the human-generated approach, people read the document, develop an understanding of its coverage and scope, apply classification and indexing rules and interpret reference sources. Objectivity and interpretation of the content may be enhanced or limited depending on the person's knowledge of the subject area, and their familiarity with rules and reference sources. Inherent time and resource limitations may result in processing lag times and delays, or it may mean that metadata can only be generated for a limited number of resources.

Hybrid Strategy

A hybrid strategy may leverage tools which perform content review, have embedded knowledge sources, emulate indexing and classification rules – behaving similar to human generation approaches. However, most of the tools that support a hybrid approach take a simplistic approach, such as:

- Exposing rules and reference sources for people to select from;
- Identifying and extracting values from embedded xml tags (i.e., author, title or date)
- Inheriting values from a business process or context or from a parent objects
- Using topic maps and making 'best guesses' based on a small set of key terms provided by catalogers and indexers;
- Using statistical clustering methods to define clusters of concepts within documents or collections.

The Bank has experimented with each of these approaches before selecting a machine-generated strategy. Our experience suggests that given a choice between human and a hybrid approach, the human-generation approach is more effective. The hybrid approach produces suboptimal results without semantic content review, capability to embed reference sources, or the capability to embed and interpret indexing and classification rules.

Machine-Generated Metadata

- A machine-generation strategy can only be effective when these capabilities are supported. This means a significant amount of up-front knowledge engineering before the technology can be deployed.

The knowledge engineering must be done up front and continuously refined until it achieves a level of quality equivalent to or better than human-generated. The critical success factors in machine-classification and indexing are the level of sophistication and flexibility of the tool. The tool needs to be able to understand and interpret the "rules".

In other words, in order to teach a machine how to classify to a "Roads and Highways" class, we need to tell it as much as we know about "Roads and Highways". This may be a set of 5,000 concepts. In order to teach a machine to identify and extract an author's name, we need to train it to recognize names, and we need to tell it where to look for names which represent an author. In both cases, we're teaching the machine to

do what a person does. In order to classify to a country, a machine needs to know what we mean by a country and how a person infers that a document is about a country. This is not the same as a simple mention of the country name, the type of output produced by the hybrid tools.

The Bank's experiences with the machine-based approach suggest:

- Metadata quality and granularity can be managed and designed, depending on the extent of design/knowledge engineering by the human architect;
- Metadata generated is richer, more comprehensive and objective than that generated by people;
- Speed of processing is significantly improved, ensuring that metadata generated within a matter of minutes after the document is produced or received;
- Tools need to be accessible to and usable by the 'knowledge engineer' (as opposed to only being exposed to the software engineer or application designer)