The Role of Descriptive Metadata & Controlled Vocabularies in Improving Semantic Linkages

STM e-Production Seminar 2014

Narendra Venkatramani (Venky)
Senior Vice President, Sales & Marketing
Delivery Channels Have Become Complex

Enriched Metadata has become key for discoverability
Types of Metadata

**Bibliographic Metadata**
- Title
- Author
- Other Identifiers (DOI, ISBN, Volume, etc.)

**Descriptive Metadata**
- Subject Classification
- Subject Keywords
- Abstracts / Summaries
- Taxonomy / Thesaurus

**Machine-aided human processing**

**Semantic Metadata**
- Content Intelligence
  - Concepts
  - Entities
  - Semantic Linkages
  - RDF & Triples

**Increasing User Experience and Better Discoverability**

**Knowledge modelling, intelligent agents**
Role of Metadata in Discoverability

- Helps enhance discoverability from document level to content level
- Helps in consistency in content discoverability through standardized indexing processes
- Increased monetization through deep discovery and multiple downloads of documents at more granular level e.g. chapters in books, tables and figures and other supplementary information
- Content powered with intelligence to automatically generate more meaningful relationships
Industry (NISO, Nielsen) Recommendations for Enriched Metadata

NISO Recommends Content Providers to provide enriched content (Abstracts & Keywords) to Discovery Service Providers for better discoverability

NISO RP-19-2014, Open Discovery Initiative

3.2.1.3 Enriched Content

The elements in Table 3 may be provided by CPs to DSPs in total or in part for each item provided to a DSP for indexing. Inclusion of enriched content in indexes and as used for relevancy ranking greatly improves the discovery experience for users, it brings particular benefit to librarians and advanced researchers who are accustomed to controlled vocabularies. Examples are provided for each metadata element in Table 4.

Table 3: Enriched content to be provided by content providers

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing data</td>
<td>One or more keywords (from controlled or uncontrolled vocabularies) to describe the content of the item.</td>
</tr>
<tr>
<td>Full Text/Transcript</td>
<td>For text items, the entirety of the document. For audio or video content, a full transcript of the spoken content of the material. May not be relevant for all indexed content.</td>
</tr>
<tr>
<td>Abstract/Description</td>
<td>Either a text summary on the content or (for non-text materials) a description of the item.</td>
</tr>
</tbody>
</table>

Source: NISO Open Discovery Initiative

High impact on sales for the books with enhanced metadata elements (short description, long description, review and autobiography) than books with less metadata elements

This graph shows the average sales per ISBN for records holding zero to all four enhanced metadata elements – short description, long description, review and autobiography.

Source: White Paper by Nielsen – The Link Between Metadata and Sales
Definitions – Descriptive Metadata, Controlled Vocabularies (CV) & Semantic Linkages

- **Descriptive Metadata**: Describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.

- **Controlled Vocabularies (CV)**: Describes the subject and topics of documents using standardized terminologies.

- **Semantic Linkages**: Relating documents using conceptual meaning rather than keywords.
Metadata Extraction at E-Production Stage (Automatic Extraction ??)

1 Survey of renewable energy

K. Heinloth

Abstract

This document is part of Subvolume C ‘Renewable Energy’ of Volume 3 ‘Energy Technologies’ of Landolt-Börnstein Group VIII ‘Advanced Materials and Technologies’.

It contains:

1.1 Introduction: Renewable energy in the past, at present and in the future

1.2 Worldwide demand on energy and potential of renewable energy

1.3 Technologies to convert renewable primary energy

1.4 Possible utilization of the different kinds of renewable energy

1.4.1 Hydropower to produce electricity

1.4.2 Wind energy

1.4.3 Sun light

1.4.4 Biomass

1.4.5 Heat from earth interior

1.4.6 Heat from water, soil and air

1.4.7 Survey of main obstacles to increase the use of renewable energy by large

1.5 Synergy effects of extensive use of renewable energy

Metadata extracted at production stage may not help in better discoverability

Generic Keywords (extracted from full text) without weightage for Relevance

- reservoir
- pumpwater
- electric
- heat
- hydrocarbon fuel
- hydrogen
- heat pump
- generator
- pump
- discharge
- dam
- environment
- bio
- gas
- diesel
- biodiesel
- ethanol
- methanol
- conversion
- grid
- radiation
- silicon
- conversion efficiency
- collector
- storage
- residus
- crop
- combustion
- bioenergy
- gasification
- furnace
- electrolysis
- electrolyze
- photochemical
- cooling
- compression
- carbon
- dioxide
- emission
- emission
- oil
- reforming
- synthesis
- fermentation
- ferment
- pyrolysis
- photo
- alcohol
- resources
- earth
- advection
- heat capacity
- thermal conductivity
- temperature
- bore hole
- exchanger
- hydrothermal
- pollution
- regenerative
- regeneration
- wood
- reservoir
- lining
- cavern
- reflector
- reformer
- efficiency
- bio-gas
- synfuel
Metadata enrichment through informative abstracts and CV indexing helps to enhance the relevance in discoverability

CV Terms:
- BIOMASS
- Electricity production
- Electrification
- Renewable Energy Sources
- Supply & Demand

Informative Abstract

1 Survey of renewable energy, part of Landolt-Börnstein - Group VIII Advanced Materials and Technologies: Numerical Data and Functional Relationships in Science and Technology, Volume 3C1 Renewable Energy.

The chapter is a survey on renewable energy. Before industrialization, renewable energy had been the only source of energy available. At present, renewable energy can be used to provide only a rather limited amount of secondary energy. Actually, renewable energy available in total worldwide could satisfy the total worldwide demand on primary energy. Annual potential of renewable energy in units of exajoules is calculated. Renewable energy may contribute only a small fraction to satisfy the total demand on "technical" energy (not including "natural" energy in form of primary biomass to be finally converted to food), mainly using hydro power to generate electricity and burning biomass, especially leftovers and waste, to produce heat and electricity. The availability of electric power from different energy sources, its restrictions and fluctuations in time are graphically illustrated. The chapter also presents a survey of technologies to convert renewable energy into final energy demand. Illustrations also include possible utilization of the different kinds of renewable energy, such as hydro power to produce electricity, wind energy, sunlight, biomass, heat from earth interior, water, soil and air. The main obstacles to increase the use of renewable energy, synergy effects of extensive use of renewable energy are also discussed.
The Controlled Vocabulary Value Chain

- **Authority File**
  - Named entities
  - Used for metadata enrichment

- **Synonym Set/Ring**
  - List of synonyms or near synonyms that are used interchangeably for retrieval purposes

- **Simplified Controlled Vocabulary**
  - Standardized sets of terms & phrases
  - Describes a domain or subject area

- **Thesaurus**
  - Hierarchical relationships used for subject browsing & navigation

- **Taxonomy**
  - Equivalence relationships
  - Associative relationships
  - Thesaurus are used for indexing and to improve the search precision & recall

- **Ontology**
  - Semantic descriptions
  - Ontologies are used for conceptual linkages across content

**Increasing Complexity**
## Amazon Kindle Fire from a Digital Forensics Perspective

With the move toward mobile computing being the trend of this technology era it is clear that our way of life and how we deal with objects in it is changing. This swift shift from large desktop computers to inexpensive, low power applications that are easily carried in our pockets or placed next to a cup of coffee on the living room table clearly changed the way we interact with media and contact friends, colleagues and family members. This also created advancement in the field of digital forensics as with every device coming to the market, studies have been conducted to investigate the possible evidence that can be found on them. As we realize that with the comfort these devices do provide as a result of their mobility they are also providing a wealth of information about the users themselves for the same reason, hence they are really valuable source of evidence in an investigation. In this paper we will discuss one of these mobile devices which is Amazon Kindle Fire. Being a new player in the mobile computing sector there haven't been enough studies of it in the field of digital forensics regarding it. In this paper we will discuss an imaging process to acquire the data from the device then we will provide an analysis of these data and their possible sources of evidence.
Amazon Kindle Fire from a Digital Forensics Perspective

**Keywords**
- Consumer electronics
- Databases
- Electronic publishing
- Fires
- Forensics
- Mobile communication
- Smart phones

Controlled Vocabularies & Semantic Linkages – Automatic Extraction of Keywords Mapped to CV
The Term “Fires” is automatically extracted and mapped with a CV in a wrong context of ‘Forest Fires’, while the article is on an electronic device called Amazon Kindle Fire. This results in irrelevant retrieval of related articles. Proper use of CV with display of broader and narrower terms helps eliminate such out-of-context terms (Fires) and related content (Forest Fires).
Abstract helps to identify the most relevant keywords on the topics discussed in the document

Keywords for Indexing
- Digital forensics
- Mobile computing
- Mobile device
- Databases

Free Indexing (From Author or Indexer)
- Amazon
- Kindle Fire
- Source of evidence
- Imaging

Subject Terms for topical browsing
- Consumer electronics
- Mobile communication
- Forensics (Digital)
Visualization of Semantic Linkages- Better User Experience
<table>
<thead>
<tr>
<th>Title</th>
<th>Survey of renewable energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>K. Heinloth</td>
</tr>
</tbody>
</table>

### Generic & Assorted keywords

- Renewable, energy, electricity, power, energy technology, wind, hydropower, power plant, plant, HPP, fuel, water, photovoltaic, PV, solar, sun, biomass, geothermal, soil, economic aspects, hydro, river, turbine, Francis, Kaplan, Pelton, cavitation, head, thermal, tidal, TPP, Three Gorge Project, sluice, weir

### Specific & Conceptual keywords

(Specialist Audience)

### Properties

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Renewable Energy types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Three Gorges project; biomass; carbon-based transportation fuel; geothermal energy; heat pump; high pressure power plant; hydrogen production; renewable energy; renewable fuel; solar power; wind energy conversion</td>
</tr>
</tbody>
</table>
About Scope

**Scope e-Knowledge Center** – award winning global KPO provider of **Content Enrichment** and Data Services

- Over 600 Knowledge Professionals (Engineers, Scientists, Physicians, Scholars, Management Graduates, Librarians, Financial & Data Analysts)
- In-house IT team with an excellent track record in developing innovative techno-human platforms for the Publishing industry
- prOdigi™ Lab for conceptualization, fostering innovation and creating products/platforms aligned with customer demand

**Core Capabilities & Solutions**

**Content Enrichment and Discovery**
- Medical Content Services
- Abstraction and Indexing Services
- Custom Taxonomy, Ontology and Classification Services
- Semantic Enrichment
- Usage Stats Analysis and Social Media

**Data Services**
- Entity Extraction & Normalization
- Authority Files Creation
- Data Acquisition, Standardization, Normalization & Maintenance
- Business Research & Editorial Services
- Big Data Solutions

Technology-enabled knowledge partner to global enterprises

ISO 9001: 2008 (QMS) Since 2004
ISO 27001: 2005 (ISMS) Since 2005
Scope’s Techno-human Platforms

- **ConSCIse™**: A unique content abstraction solution that delivers SEO Keyword-rich abstracts to enable greater discoverability.
- **TranSCIse™**: Non-English Content…Abstracted, a technology-enabled hybrid abstracting and indexing solution for enhancing the discoverability of non-English literature in search engines and A&I services.
- **AuthEntik™**: Author Data…Enhanced, an author data management solution capable of parsing, standardizing, normalizing and disambiguating a high volume of author information from original documents.
- **InDEXr™**: A Content Indexing Solution, a platform-enabled indexing solution, offers a unique opportunity to enhance discoverability of content by providing highly-relevant indexing of key concepts/entities for a wide range of textual and non-textual documents.
- **SemantiCz™**: Concepts Linkages… Made Smarter, a platform based solution to enhance knowledge discovery through linked data of concepts and relationships using Ontologies and RDF.
- **procuRx™**: Spend Data… Analyzed, a solution designed to provide and enhance the visibility of an organization’s spend.
- **OrdEHRSet™**: Hospital Information System…Enabled, offers faster, accurate and high-quality order sets that enable a healthcare information system to quickly reach maximum functionality and usability in meeting clinicians’ needs.
- **diSCOveR™**: Your Content…More Discoverable, a technology-enabled hybrid abstracting and indexing solution for enhancing the discoverability of non-English literature in search engines and A&I services.
- **mARCat™**: Meta Data…Catalogued, a platform-based solution to create MARC records for input files with dynamic cataloging template based on MARC standards, which will be validated by subject matter experts (SMEs).

**Techno-human Model**
- Service as a Product
- Scalable & Robust
- Non-linear Business Model
- Faster Go-To-Market
- Enhance Market Share

**Platforms**
- Non Linearity
- Scale

**Products for Publishers**
- **prodigi™**

**Products for Non Publishers (enterprises, portals, digital libraries)**
- **OrdEHRSet™**
- **diSCOveR™**
- **mARCat™**
Thank You

www.scopeknowledge.com
www.knowledgespeak.com

Narendra Venkatramani (Venky)
Senior Vice President, Sales & Marketing
venky@scopeknowledge.com