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

Frankfurt Book Fair 2014

Rich Kobel
Vice President,
Business Development
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 and Social Media Analytics

#### Data Services
- Entity Extraction & Normalization
- Authority Files Creation
- Data Acquisition, Standardization, Normalization & Maintenance
- Big Data Solutions
- Business Research & Predictive Analytics

---

**Technology-enabled knowledge partner to global enterprises**

**ISO 9001: 2008 (QMS) Since 2004**

**ISO 27001: 2005 (ISMS) Since 2005**

---

**Content Enrichment and Discovery**

<table>
<thead>
<tr>
<th><strong>Medical Content Services</strong></th>
<th><strong>Abstraction and Indexing Services</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Custom Taxonomy, Ontology and Classification Services</strong></td>
<td><strong>Semantic Enrichment</strong></td>
</tr>
<tr>
<td><strong>Usage and Social Media Analytics</strong></td>
<td></td>
</tr>
</tbody>
</table>

---

**Data Services**

- Entity Extraction & Normalization
- Authority Files Creation
- Data Acquisition, Standardization, Normalization & Maintenance
- Big Data Solutions
- Business Research & Predictive Analytics

---

**Deep Subject Matter Expertise**

**IT-Enabled Platforms**

**Multi-Lingual Capabilities**
Types of Metadata

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

Descriptive Metadata

Content Structuring
- Subject Classification
- Subject Keywords
- Abstracts / Summaries
- Taxonomy / Thesaurus

Semantic Metadata

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

Knowledge modelling, intelligent agents

Increasing User Experience
And Better Discoverability
Publishing Supply Chains Have Become Complex

Enriched Metadata has become key for discoverability
Role of Metadata in Discoverability

- Helps to enhance discoverability from document level to content level
- Helps in consistency in content discoverability through standardized indexing process
- 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, Neilson) 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>

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. Clearly see the high impact on sales that having a data rich product record has.

Source: NISO Open Discovery Initiative

Source: White Paper by Neilson – 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.
Examples of Descriptive Metadata & Controlled Vocabulary Indexing

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

CV helps to indicate the “aboutness” of the content by assigning subject terms across corpus of documents.
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

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

- **Thesaurus**
  - List of synonyms or near synonyms that are used interchangeably for retrieval purposes

- **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
Semantic Linkages – The Final Lap to Knowledge Discovery

- **Silicon**
  - Has property: Periodic Lattice
  - Short-Range Order
  - Has sub-class: Amorphous Silicon
  - Crystalline Silicon
  - Has type: Single-Crystalline Silicon
  - Polycrystalline Silicon
  - Thin-Film Silicon
    - Used in: Thin-Film Transistors
    - Passivation Layers
    - Solar Cells
      - Has type: Single-Junction Solar Cells

- **Silicon Compounds**
  - Has sub-class: Silane
    - RF-Driven Glow Discharge
  - Plasma-Enhanced Chemical Vapour Deposition
  - Hot-Wire Chemical Vapour Deposition
  - Catalytic Hot Surface
    - Using: Silane
    - Decomposition
    - Undergoes: Silane

- **Silicon-Containing Gases**
  - Has sub-class: Silane
    - Forming: Plasma-Enhanced Chemical Vapour Deposition
    - Forming: Hot-Wire Chemical Vapour Deposition

---

Quattro Confidential

© Copyright Scope e-Knowledge Center 2014
Visualization of Semantic Linkages- Better User Experience
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

<table>
<thead>
<tr>
<th>Abstract</th>
<th>Authors</th>
<th>References</th>
<th>Cited By</th>
<th>Keywords</th>
<th>Metrics</th>
<th>Similar</th>
</tr>
</thead>
</table>

#### Subject Terms
- Consumer electronics
- Databases
- Electronic publishing
- Fires
- Forensics
- Mobile communication
- Smart phones
The Term “Fires” is automatically extracted and mapped with a CV in a wrong context of ‘Forest Fires’, while the article is on a 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

Semantic Fingerprint
Keywords can also be conceptually related to provide a semantic fingerprint of the document and retrieve more relevant related documents

Kindle fire manufactured by Amazon is a mobile device used for mobile computing helps in generating databases extracted through an imaging process as source of evidence in digital forensics
# How Metadata is Enriched for Different Users of the Same Content

<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th><strong>Survey of renewable energy</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author</strong></td>
<td><strong>K. Heinloth</strong></td>
</tr>
</tbody>
</table>

**Generic & Assorted keywords**

<table>
<thead>
<tr>
<th>(General Audience)</th>
</tr>
</thead>
<tbody>
<tr>
<td>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</td>
</tr>
</tbody>
</table>

**Specific & Conceptual keywords**

<table>
<thead>
<tr>
<th>(Specialist Audience – Material Scientists)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy type</td>
</tr>
<tr>
<td>Three Gorges project; biomass; carbon-based transportation fuel; geothermal energy; heat pump; high pressure power plant; hydroelectric power; hydrogen production; renewable energy; renewable fuel; solar power; wind energy conversion</td>
</tr>
</tbody>
</table>
Scope releases the enhanced version of SemantiCz™, its semantic enrichment solution for better discovery. The key feature & functionalities of enhanced version SemantiCz 2.0 include:

- Advanced text mining and ontology based semantic tagging
- Generate relationships with precise, subject-specific contextual accuracy
- Manage multiple versions of domain specific Ontologies, in various languages
- Extract relationships from the content, specifically predicates, using advanced text mining algorithms thereby providing a high level of accuracy
- Integrate with external platforms like MarkLogic, to ensure that the enriched information plugs in seamlessly with the client’s ecosystem.
- Handle a wide range of subject domains based on its proprietary engine and Scope’s vast experience in building taxonomies in the STM domains
**Techno-human Platforms**

**ConSCIse™**
A Unique Content Abstraction Solution
ConSCIse™, a unique content abstraction solution that delivers SEO Keyword-rich abstracts to enable greater discoverability

**TranSCIse™**
Non-English Content...Abstracted
TranSCIse™, 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
AuthEntik™, 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
InDEXr™, 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
SemantiCz™, a platform based solution to enhance knowledge discovery through linked data of concepts and relationships using Ontologies and RDF

**prodiCgi™**
Platforms ● Non Linearity ● Scale

**procuRx™**
Spend Data... Analyzed
procuRx™, a solution designed to provide and enhance the visibility of an organization’s spend

**OrdEHRSet™**
Hospital Information System...Enabled
OrdEHRSet™ 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
diSCOver™, packages robust content mining, semantic enrichment and controlled vocabulary building and maintenance capabilities to improve discoverability and searchability

**mARCat™**
Meta Data...Catalogued
mARCat™, 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 service platforms to enhance discoverability, enrich content and interpret knowledge**

- Products for Publishers
- Products for Non Publishers (enterprises, portals, digital libraries)
Thank You

www.scopeknowledge.com
www.knowledgespeak.com

Rich Kobel
Vice President
Business Development
rkobel@scopeknowledge.com
+1 516 462 3555