

Artificial Intelligence Secret Ingredient? Information Architecture

A.J. Rhem

 & Associates Inc.™

Providing Today's Solutions
with Tomorrow's Technology

World IA Day 2019 Kent State University



About Your Presenter

Dr. Anthony J. Rhem, PhD.: serves as the President and Principle Consultant of **A.J. Rhem & Associates, Inc.**, a privately held Knowledge Management & System Integration Consulting, Training and Research firm located in Chicago, Illinois.

Dr. Rhem has over thirty (30) years of experience in information technology and twenty years (20) in Knowledge Management. A published author, educator, and researcher; Dr. Rhem has presented the application and theory of Software Engineering Methodologies, Knowledge Management, Artificial Intelligence, Information Architecture, Big Data and IoT at universities and conferences in the US, Europe and Australia.

Introduction

Today's discussion on Artificial Intelligence (AI) is primarily focused on Machine Learning, Cognitive Computing and Big Data Analytics.

However, you must prepare your organization's data in order to properly take advantage of AI tools that are focused on Big Data Analytics (such as, [IBM](#), [Amazon](#), [Microsoft](#), and [Google](#)).

To properly prepare your data you will need to apply Information Architecture.

Why Information Architecture (IA)?

- IA provides the process, procedures and methods to perform Data Preparation
- IA focuses on the semi-structured and unstructured data that comprises over 90% of the data being analyzed by big data analytics.
- Semi-structured data: is a form of data that does not conform with the formal structure of data in a databases or data tables, but contains tags to separate elements and enforce hierarchies within the data (i.e., spreadsheets, XML files).
- Unstructured data: is a form of data with no tagging, metadata or inherent structured associated with it (i.e., image, text, voice, video). Content typically refers to the container that the semi-structured and unstructured data resides in (i.e., .pdf, .doc, .xml, .ppt, .csv).

Tenets of Information Architecture for AI Data Preparation

Information Architecture Focus:

Information Architecture: Information architecture (IA) focuses on **organizing**, **structuring**, and **labeling** content in an effective and sustainable way. The goal is to help users find information and complete tasks. To do this, you need to understand how the **pieces fit together to create the larger picture**, how items relate to each other within the system – **Usability.gov**

Information Architecture



Classifications & Hierarchies



Labels & Indexing



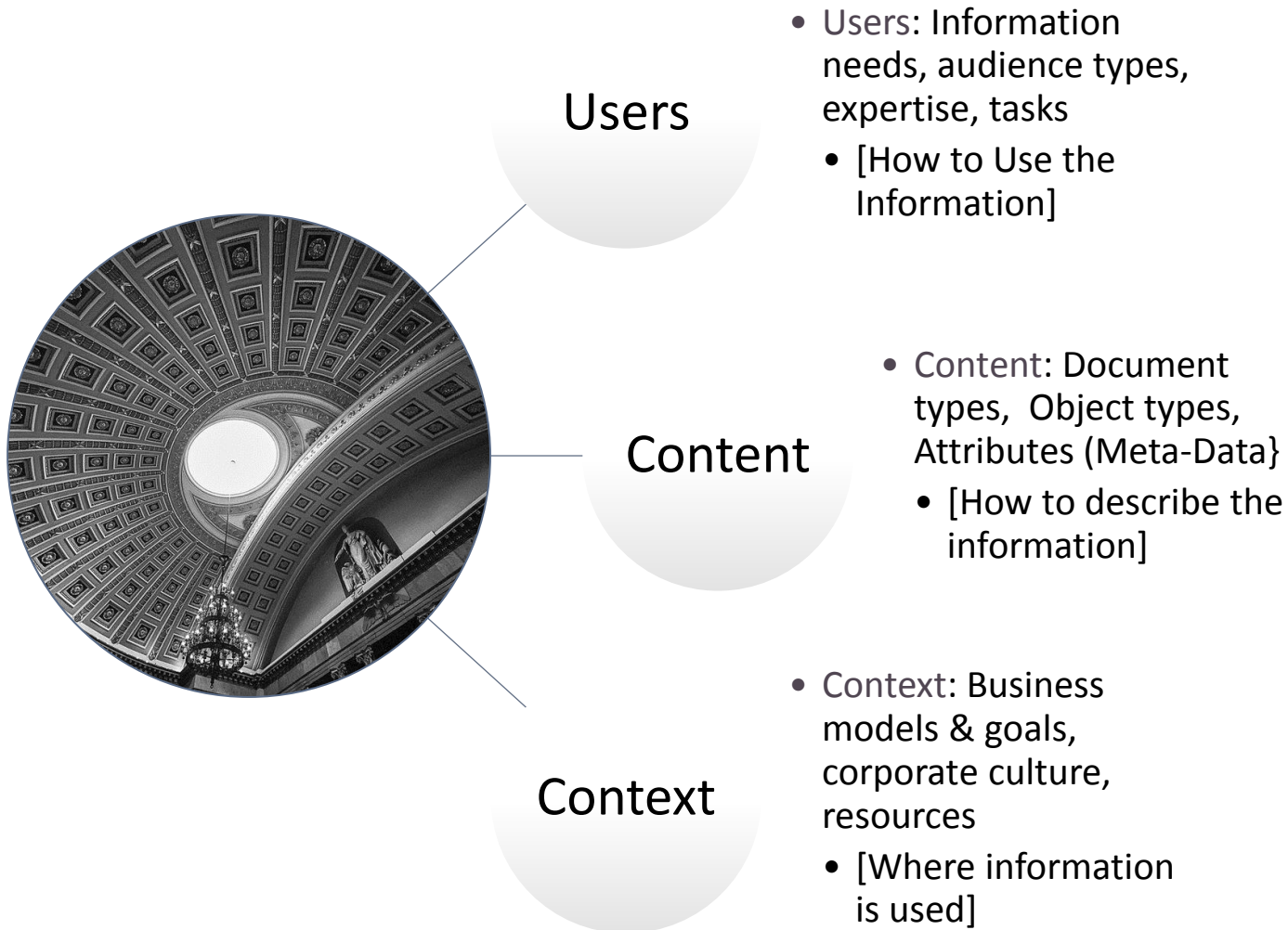
Navigation



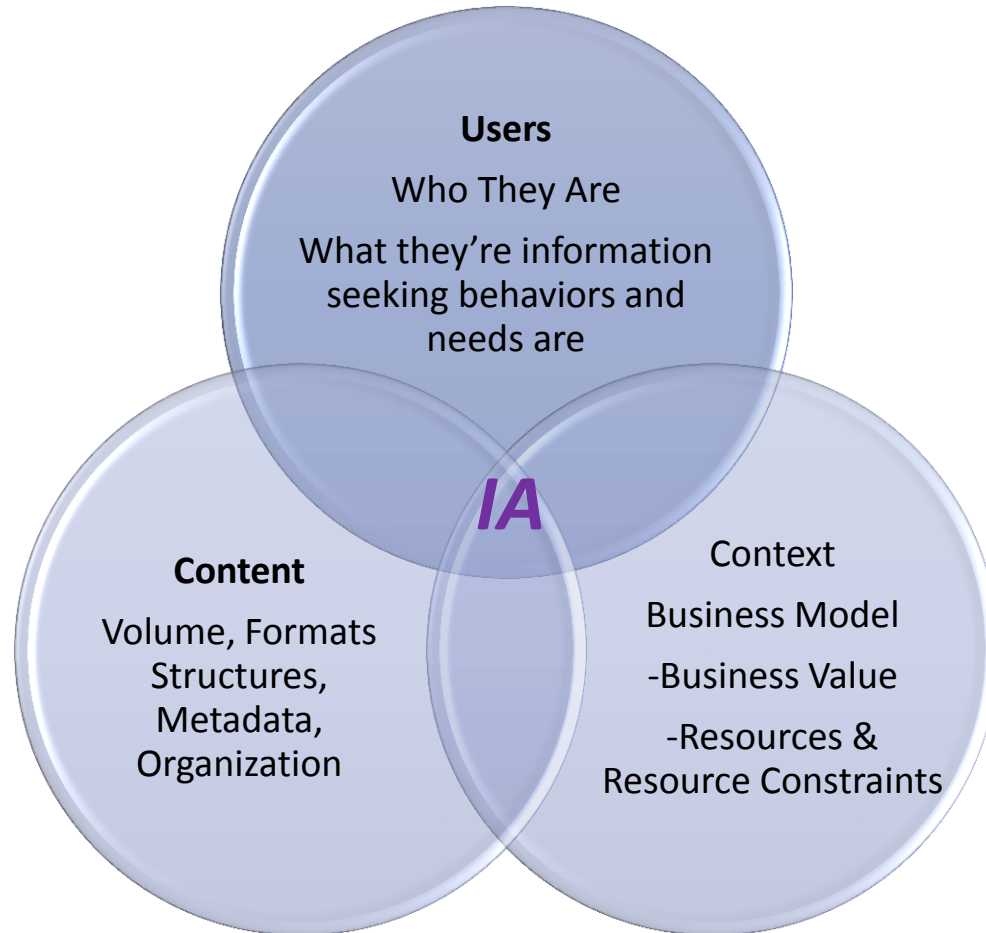
Search

INFORMATION ARCHITECTURE CONNECTS PEOPLE TO CONTENT

Three Major Parts of Information Architecture



Information Architecture Areas of Practice





Data Preparation

Using Information Architecture

Data Preparation

- Data preparation provides the methodological and technological data management support to address data quality issues, maximize the usability of the data; provide an active and on-going management of data through its lifecycle; perform data discovery and retrieval, create and maintain quality, add value, and provide for re-use over time.

Data preparation process includes the following activities:

- Data Audit
 - Determine what data is ready to be consumed, evaluate the quality of the data
 - Determine the gaps in the data needed, and identify the criteria to determine what data is used (and not used).
- Data Analysis
 - Data analysis examines information concepts, relationships, business rules and metadata. This provides a sharable, stable and organized structure for the data under consideration.

Data preparation process includes the following activities:

- Address Data Gaps
 - The results of performing a Data Audit will determine the gaps in data and identify the additional sources of data that are needed for effective data analytics.
- Data Selection & Validation
 - Data selection should be considered in terms of significance, how essential or basic is it to the discipline; validity, is the data accurate, current and relevant to the domain under consideration.

Data preparation process includes the following activities:

- Classification
 - The classification of data will be in the form of one or more ontologies/taxonomies. Classification of data will also be realized through controlled vocabularies and thesaurus. The structure refers to the methods to aggregate the concepts and metadata into the domain ontology/taxonomy.
- Align Data to Domain Ontology/Taxonomy
 - Categorizing data and aligning the data to a common ontology/taxonomy is essential to big data analytics due to the varied number of data sources under consideration.
- Transformation
 - Transformation provides consistency identified by standard and precise metadata; aligned to an accurate and exact ontology/taxonomy

Case Study: Cancer Research Tool

The Sherlock™ Big Data Knowledge Discovery & Analysis tool is an application that will use AI Machine Learning Algorithms to “mine” Big Data Cancer repositories to extract knowledge. This knowledge will be used to assist oncologists and cancer researchers in enhancing diagnostic decision making for patients with Non-Small Cell Lung Cancer (NSCLC) and in discovering new treatment strategies.



NSCLC
Ontology



Auto-
classification
(Information
Architecture)



Knowledge
(Tacit-Explicit)
Acquisition



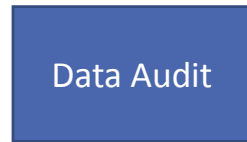
Self-learning Neural
Network

Machine Learning Data Preparation Process Using IA

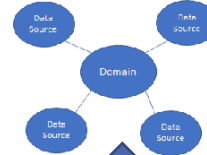
Data from Various Relevant Sources are identified



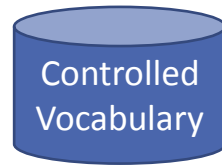
Determine what data is ready to be consumed and fill the gaps in the data



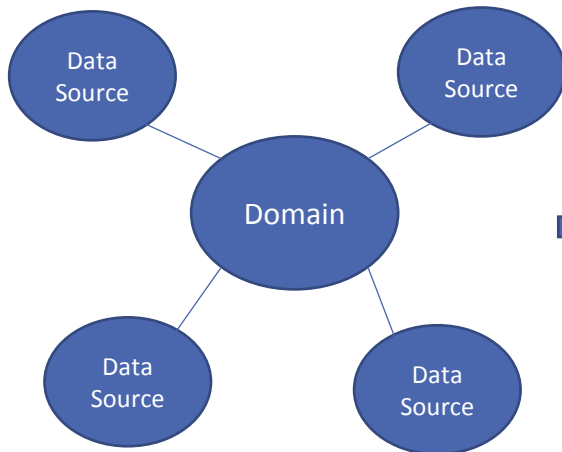
Construct Common ontology; Tag semi and unstructured data based on ontology terms



Ontology terms are transitioned to Controlled Vocabulary for Labeling and Indexing



Map Data to a Common Ontology



Transformation of the data has occurred: consistently identified with standard and precise metadata; map to a common ontology based on the domain under analysis



Determine and model Algorithm

Data Discovery: Cancer Research Tool

Started with Specific Questions to be Answered by the data:

- What is the Health Trajectory of patients with NSCLS (Health Trajectory Analysis - build a finite state model)
- What treatments are most effective for NSCLC at the various stages?
- What specific genomic and environmental factors contribute to NSCLC onset and the effect of treatment options
- How can the data lead to detecting NSCLC early?

Data Selection:

- National Cancer Database (NCDB)
- National Cancer Institute (NCI) Data Catalog
- U.S. National Library of Medicine Clinical Trails Database

National Cancer Database (NCDB):

A clinical oncology database sourced from hospital registry data that are collected in more than 1,500 Commission on Cancer (CoC)-accredited facilities. NCDB data are used to analyze and track patients with malignant neoplastic diseases, their treatments, and outcomes. Data represent more than 70 percent of newly diagnosed cancer cases nationwide and more than 34 million historical records.

National Cancer Database Sample Data

Lung Cancer Mortality Rates	Lung Cancer Incident Rates	Median Family Income	Level of Education (High School or >)	State
43.1%	59.6%	\$67,023	90.05%	Nebraska
48.1%	67.1%	\$77,283	79.85%	Illinois
51.1%	71.5%	\$87,019	82.15%	New York
49.1%	62.8%	\$72,001	70.30%	California
33.7%	49.6%	\$91,155	78.42%	Texas

National Cancer Institute (NCI) Data Catalog:

The NCI Data Catalog is a consolidated listing of the publicly available data collections produced by NCI initiatives, including The Cancer Genome Atlas (TCGA), The Cancer Imaging Archive (TCIA) and Surveillance, Epidemiology, and End Results (SEER) database. Contains a variety of data (Structured, Semi-structured and Unstructured)

National Cancer Institute (NCI) Data Catalog: Sample Data

Variable accession	Variable name	Variable description
phv00088007.v2.p1	SUBJID	CGCI Subject Id
phv00088008.v2.p1	WHO diagnosis	WHO diagnosis
phv00088009.v2.p1	Days to birth	Number of days to birth
phv00088010.v2.p1	Gender	Gender
phv00088011.v2.p1	Stage	Stage (Ann Arbor)
phv00088012.v2.p1	Stage group	Stage (limited vs advanced)
phv00088013.v2.p1	Performance status	Performance status (ECOG)

U.S. National Library of Medicine Clinical Trails Database (PubMed):

PubMed comprises more than 29 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites. (This data is primarily Unstructured Data)

U.S. National Library of Medicine Clinical Trails Database Sample Data

[Prospective Multicentered Safety and Feasibility Pilot for Endobronchial Intratumoral Chemotherapy.](#)

1. Yarmus L, Mallow C, Akulian J, Lin CT, Ettinger D, Hales R, Voong KR, Lee H, Feller-Kopman D, Semaan R, Seward K, Wahidi MM.
Chest. 2019 Feb 15. pii: S0012-3692(19)30155-2. doi: 10.1016/j.chest.2019.02.006. [Epub ahead of print]
PMID: 30776363
[Similar articles](#)

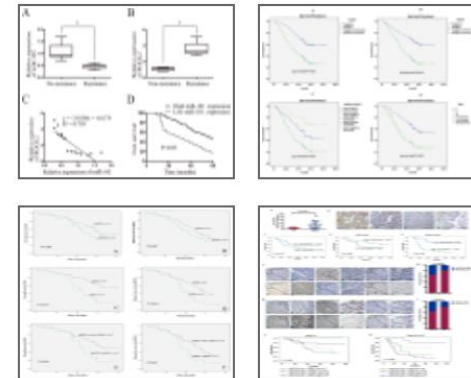
[Safety and effectiveness of alectinib in a real-world surveillance study in patients with ALK-positive NSCLC in Japan.](#)

2. Noriyuki M, Ohe Y, Gemma A, Kusumoto M, Yamada I, Ishii T, Yamamoto N.
Cancer Sci. 2019 Feb 18. doi: 10.1111/cas.13977. [Epub ahead of print]
PMID: 30776174 **Free Article**
[Similar articles](#)

[Lung cancer family history and exposure to occupational/domestic coal combustion contribute to variations in clinicopathologic features and gene fusion patterns in non-small cell lung cancer.](#)

3. Chen Y, Li G, Lei Y, Yang K, Niu H, Zhao J, He R, Ning H, Huang Q, Zhou Q, Huang Y.
Thorac Cancer. 2019 Feb 18. doi: 10.1111/1759-7714.12987. [Epub ahead of print]
PMID: 30775858
[Similar articles](#)

PMC Images search for NSCLC



See more (17467)...

Data Audit: Cancer Research Tool

- The data audit is based on the ability of the datasets under consideration to answer the data analysis questions:
 - What is the Health Trajectory of patients with NSCLS (Health Trajectory Analysis - build a finite state model)
 - What treatments are most effective for NSCLC at the various stages?
 - What specific genomic and environmental factors contribute to NSCLC onset and the effect of treatment options
 - How can the data lead to detecting NSCLC early?
- The data audit will: determine what data is ready to be consumed; evaluate the quality of the data, determine the gaps in the data needed, and identify the criteria to determine what data is used (and not used) from the selected datasets:
 - National Cancer Database (NCDB)
 - National Cancer Institute (NCI) Data Catalog
 - U.S. National Library of Medicine Clinical Trails Database

Data Analysis: Cancer Research Tool

The Data analysis examined data concepts, relationships, business rules and metadata. contributing to:

- Development the NSCLC Ontology
- Determining controlled vocabulary
- Selection of metadata

Data Classification/Standardization

NSCLC Ontology Structure

- Our product ontology is based on the NSCLC disease patterns as determined by the data from the National Center for Biomedical Ontology.

It is a combination of the following ontology resources:

[NSLC Biomedical Ontology - Scientific Projects](#)

[NSCLC Physicians Data Query](#)

[Human Disease Ontology](#)

[Interlinking Ontology for Biological Concepts](#)

Cancer Research Tool NSCLC Ontology Structure

[NSCLC Ontology Viewer](#)

NSCLC ML Algorithm

- A deep learning supervised algorithm was used to...
 - This algorithm will facilitate predictions on given set of samples.
 - Searches for patterns within the value labels assigned to data points.
 - This algorithm used labeled training set that contains both normal and anomalous samples for constructing the predictive model.
 - Based on Bayes Probability Theorem

Conclusion

- Defining a structured representation associated with the data allows users to compare, aggregate, and **transform the data**.
- With greater data availability, **the barrier of data acquisition is reduced**. To **extract value from the data it needs to be systematically processed, transformed, and repurposed into a new context**.
- Data Preparation of semi-structured and unstructured data in big data analytics is driven by the need to **reduce the time-to-market, reduce the time to create new products, repurposing existing content and to improve accessibility and visibility of information artifacts**.
- Data Preparation is important because of the growth of the variety of sources used in Big Data. Selecting data from a variety of well curated sources will add richness to the Big Data Analytics.

THANK YOU!



WIAD 2019 KENT, OHIO
LOCAL CONNECTIONS. GLOBAL IMPACT.



Want Better Products and More Business Value? Design Better Organizations!

Anthony J. Rhem, PhD
Principal, A.J. Rhem & Associates, Inc.
@AJRhemAssoc | @WIADkent



THE INFORMATION ARCHITECTURE INSTITUTE

Learn more at: <https://2019.worlddiaday.org>

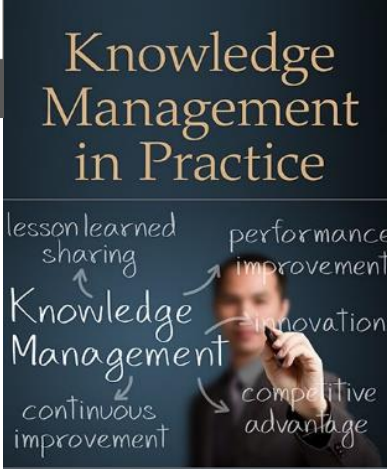
email: tonyr@ajrhem.com

Website: www.ajrhem.com

LinkedIn: www.linkedin.com/in/anthonyrhem


Blog: <http://knowledgemanagementdepot.com/>

Latest Book:



Knowledge Management in Practice

Anthony J. Rhem



CRC Press
Taylor & Francis Group
AN AUERBACH BOOK