Indexing in a Digital World

Daniel Gelaw Alemneh
University of North Texas, USA
Fulbright Scholar at University of Pretoria, South Africa

ASSOCIATION OF SOUTHERN AFRICAN INDEXERS AND BIBLIOGRAPHERS (ASAIB)

Indexing in Service of the Reader

ASAIB PRE-CONFERENCE WORKSHOPS & ANNUAL CONFERENCE 2022

19 – 20 May 2022

Introduction

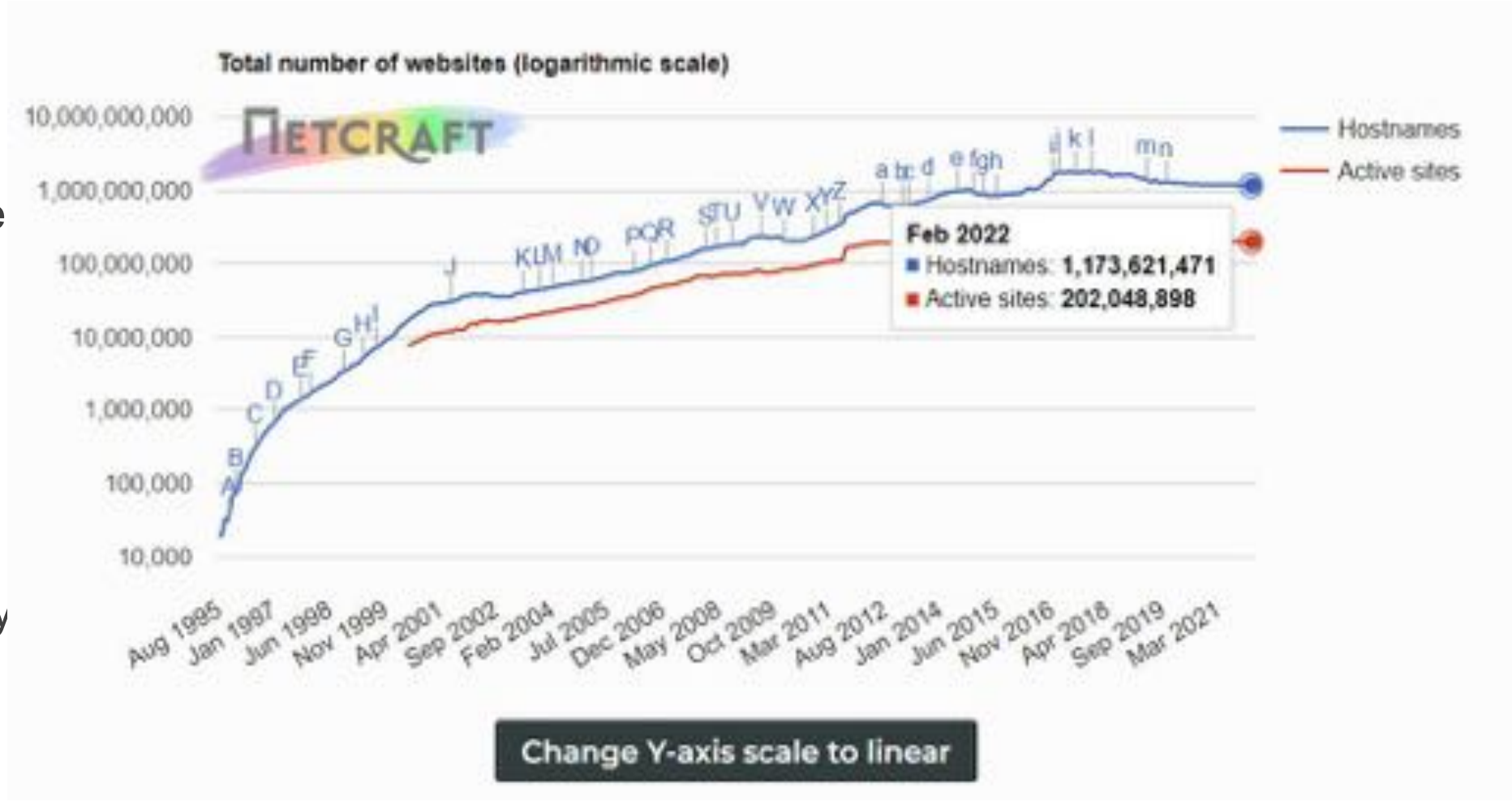
- The synergies of numerous emerging trends (advancement in information communication technologies, cross discipline collaborations, development of open standards and open-source software, etc.) are making resource integration much easier and provide users with access to more diverse and previously unavailable contents and services that span myriad technologies across institutions and nations.

- The diverse, huge, and ever-expanding digital information resources in the Web, has evolved without much regard to resources management and organization issues.

Introduction

- According to the most recent (February 2022) Web Server Survey , they received response from 1,2 billion sites across 271,199,972 unique domains and 11,774,714 web-facing computers.

- No wonder our capacity to manage and digest this information explosion has not kept up.



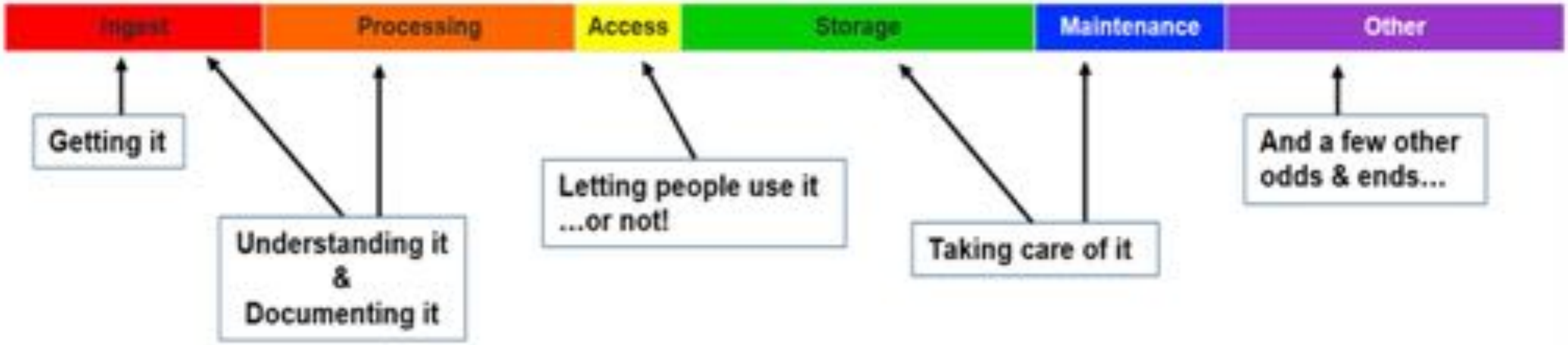
Source: Netcraft Web Server Survey: (Links to an external site.)https://news.netcraft.com/archives/category/web-server-survey/

Ongoing Challenges

- Although new technologies have been developed for data storage, data volumes are doubling in size about every two years.
  - The Cisco survey estimates that in 2018, the total amount of global data storage capacity to grow to an estimated 1.1 zettabytes (ZB) , which is approximately twice the space available in 2017 (600 Exabyte (EB).
  - Annual global data center IP traffic reached 15.3 ZB (1.3 ZB per month) by the end of 2020.
- Curating and preparing data before it can actually be used.

Digital Curation Workflows Considerations

- **Digital curation is the management, preservation, and enrichment of digital resources.**



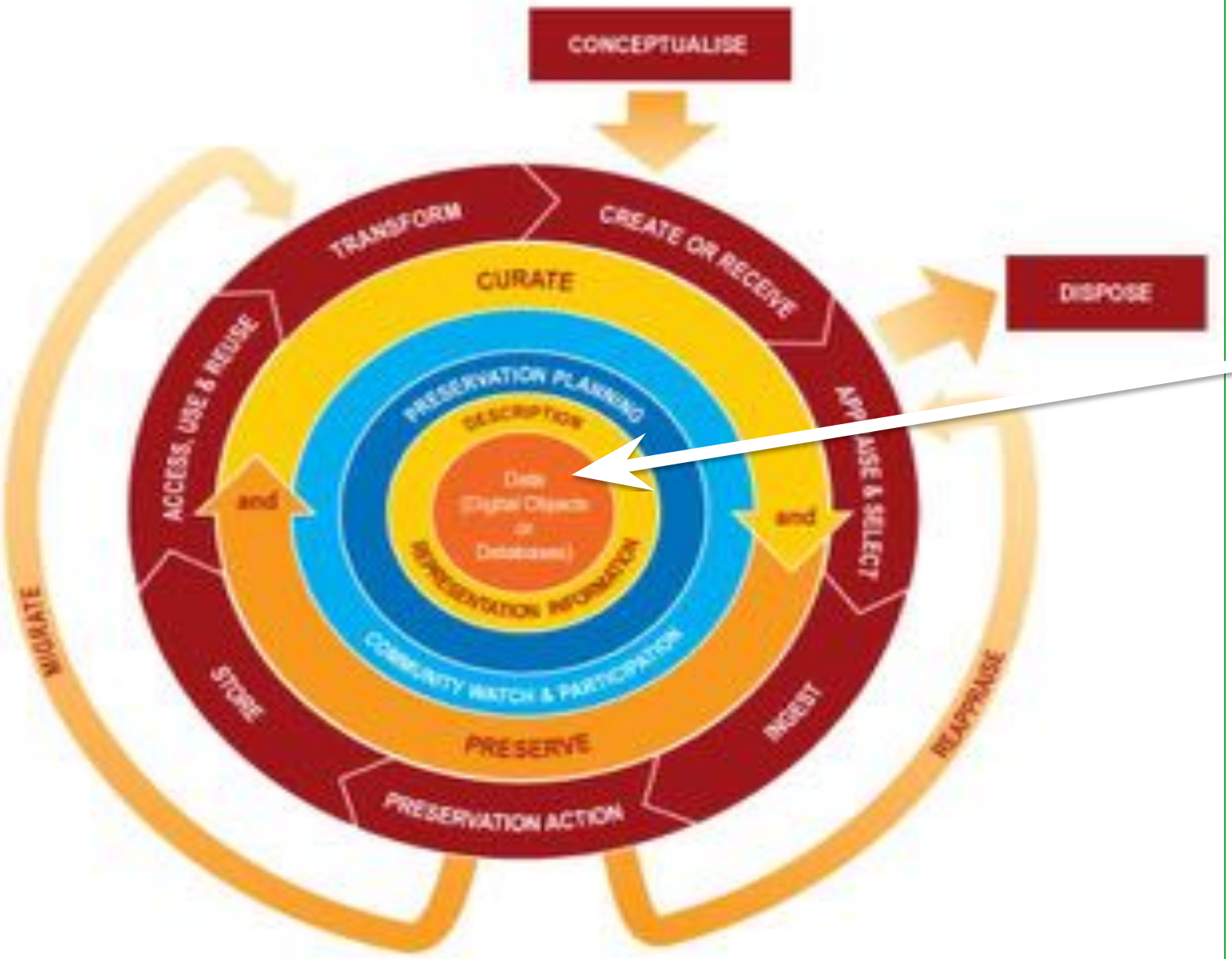
Source: A Digital POWRR Workshop: http://digitalpowrr.niu.edu/

Digital Curation Workflows Considerations

- **Various tools support the processes of creating, organizing, discovering, retrieving, using, and preserving information.**

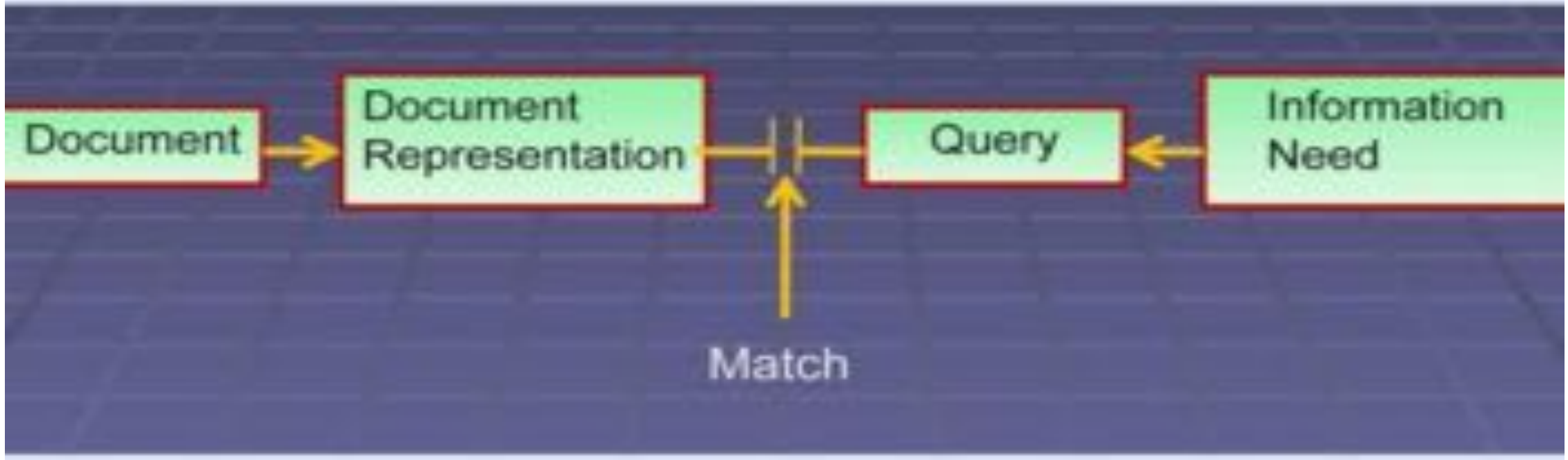
Source: A Digital POWRR Workshop: http://digitalpowrr.niu.edu/

Indexing in Data Curation Lifecycle



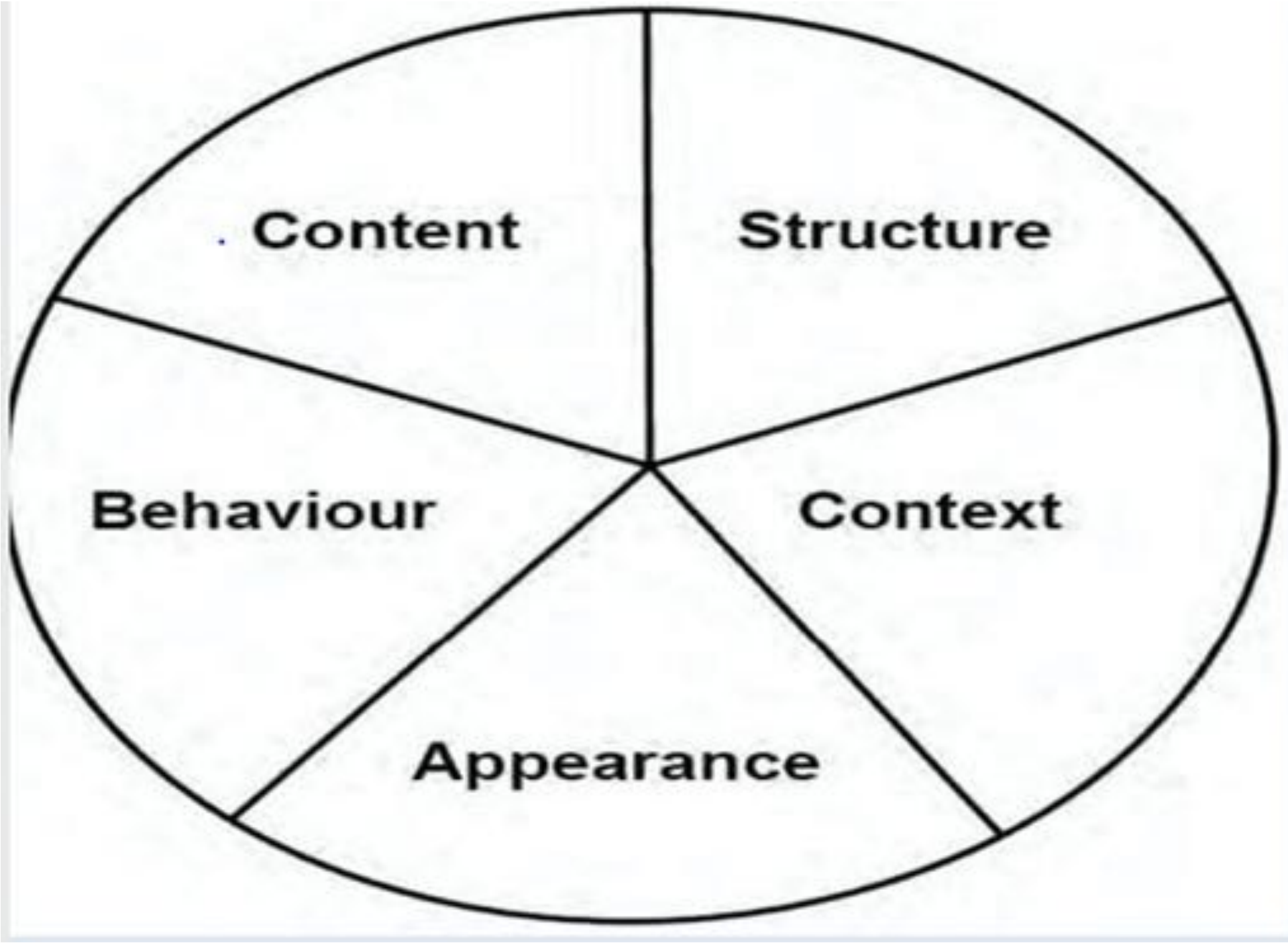
Indexing and Metadata for discovery and reuse

The Classic Information Retrieval (IR) Model



Source: (Modified from) Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407-424.

Five Basic Aspects of Digital Object



Aboutness

- Fundamentally, non-print records of knowledge are means of expression and communication, which transcend language and words.
  - “Aboutness" never has the concept been more pertinent than when we get into nonverbal information.

- Despite the promise and sophistication of content based image indexing and retrieval (CBIIR) tools, human assignment of indexing terms and description still play significant role in image retrieval.

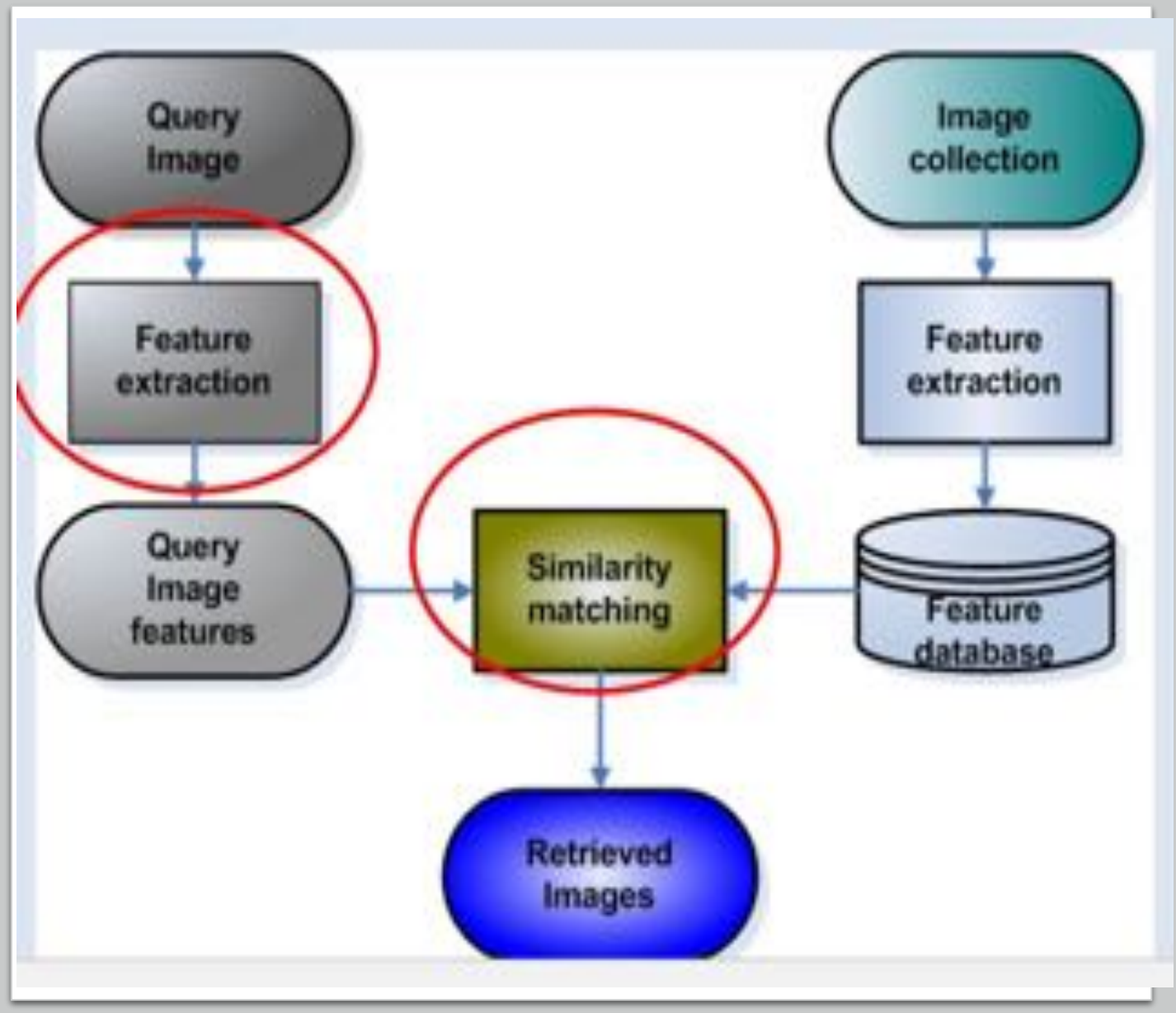
[*Aboutness*] is in the Eye of the Beholder

- **What does the painting of Mona Lisa mean to:**
  - an eight-year-old taking a class tour through the Louvre
  - an art historian from South Africa
  - a long-jawed art student from UP
  - a tourist from Canada
  - an art thief from a big city,
  - an indexer or attendee of the ASAIB-2022 Conference

Source: https://www.wikiart.org/en/leonardo-da-vinci/mona-lisa

A picture is worth a thousand words…

- But how do we know when exhaustivity has gotten to a degree where the index is no longer being improved?



Rorissa, A. (2007). Benchmarking Visual Information Indexing and Retrieval.. ASIS&T Bulletin, February/March 2007. Retrieved from: https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/bult.2007.BULT1720330310

Bridging the Gap Between Abstracting & Indexing Provider Needs and Discovery Service Approaches

Abstract

A report from the NISO Open Discovery Initiative Standing Committee. Authors: Scott Bernier, EBSCO Information Services; Noah Brubaker, PALNI; Rachel Kessler, ProQuest; Geoffrey Morse, Northwestern University; Ken Varnum, University of Michigan.

In 2014, NISO published a Recommended Practice, *Open Discovery Initiative: Promoting Transparency in Discovery* (RP-19-2014), developed by its Open Discovery Initiative (ODI) Working Group to define "standards and/or best practices for the new generation of library discovery services that are based on indexed search." In addition to recommending specific practices for discovery and content providers to adopt, the document recommended a number of potential future work items. One of these was that the ODI initiative should take an in-depth look at the unique needs of Abstracting & Indexing (A&I) service providers with regard to the inclusion of their content in web scale discovery systems. The goal of such an assessment was to "identify subsequent recommendations that would address these concerns and ultimately encourage discovery participation from providers of these services." Section 2.2 of the Recommended Practice specifies several concerns expressed by A&I service providers regarding discovery systems. These concerns included:

- The level of exposure A&I citations will receive in discovery services;
- The possibility that the presence of discovery systems will encourage libraries to cancel A&I subscriptions; and
- The probability that discovery services would not be able to take advantage of the sophisticated controlled vocabulary and other features that A&I service providers offer in their native interfaces.

In June 2017, a subgroup of the Open Discovery Initiative Standing Committee conducted a survey of A&I providers to understand their concerns. This was followed in February 2018 with a related survey of discovery providers. This report describes these two related survey efforts, summarizes concerns raised by both groups, and recommends next steps for the Open Discovery Initiative Standing Committee to better promote transparency and understanding among discovery providers, abstracting and indexing providers, and librarians.

Publication type

Other

Front Matter

Publication Date: June 21, 2019

ANSI/NISO Z39.4-2021 Criteria for Indexes

Abstract

This standard provides guidelines for the content, organization, and presentation of indexes used for the retrieval of documents and parts of documents. It deals with the principles of indexing regardless of the type of material indexed, the indexing method used, the medium of the index, or the method of presentation for searching. It emphasizes three processes essential for all indexes: comprehensive design, vocabulary management, and syntax.

An American National Standard
Developed by the National Information Standards Organization

Approved: July 12, 2021 by the American National Standards Institute

Published by the National Information Standards Organization
Baltimore, Maryland, U.S.A.

Description

See the Z39.4 Criteria for Indexes Working Group page

Publication type

Standard

Front Matter

Publication Date: July 14, 2021
ISBN: 978-1-950980-14-7
DOI: 10.3789/ansi.niso.z39.4-2021
ISSN: 1041-5653

ANSI/NISO Z39.4-2021 Criteria for Indexes

- This standard provides guidelines for the content, organization, and presentation of indexes used for the retrieval of documents and parts of documents.

- It deals with the principles of indexing regardless of:
  - the type of material indexed,
  - the indexing method used,
  - the medium of the index, or
    - the method of presentation for searching.

ANSI/NISO Z39.4-2021 Criteria for Indexes...

- It emphasizes three processes essential for all indexes:
  1. Comprehensive design,
  2. Vocabulary management, and
  3. Syntax.

Most Important Metadata Quality Criteria

(Park & Tosaka, 2010):

- Access
- **Accuracy**
- Availability
- Compactness
- Compatibility
- **Completeness**
- Comprehensiveness
- Content
- **Consistency**
- Cost
- Data structure
- Ease of creation
- Ease of Use
- Economy
- Flexibility
- Fitness for Use
- Informativeness
- Protocols
- Quantity
- Reliability
- Standard
- Timeliness
- Transfer
- Usability

Most Important Metadata Quality Criteria

(Park & Tosaka, 2010):

Accuracy

- *Format and formatting errors*
- *Spelling and typographical errors*
- *Have accepted methods been used for creation or extraction of metadata?*
- *What has been done to ensure valid values and structure?*
- *Are default values appropriate, and have they been appropriately used?*

Completeness

- *Number of elements per record*
- *Practice of presenting "blank" elements*
- *Utilization and selected characteristics of "mandatory" and "optional" elements*
- *Does the element set completely describe the objects?*
- *Are all relevant elements used for each object?*

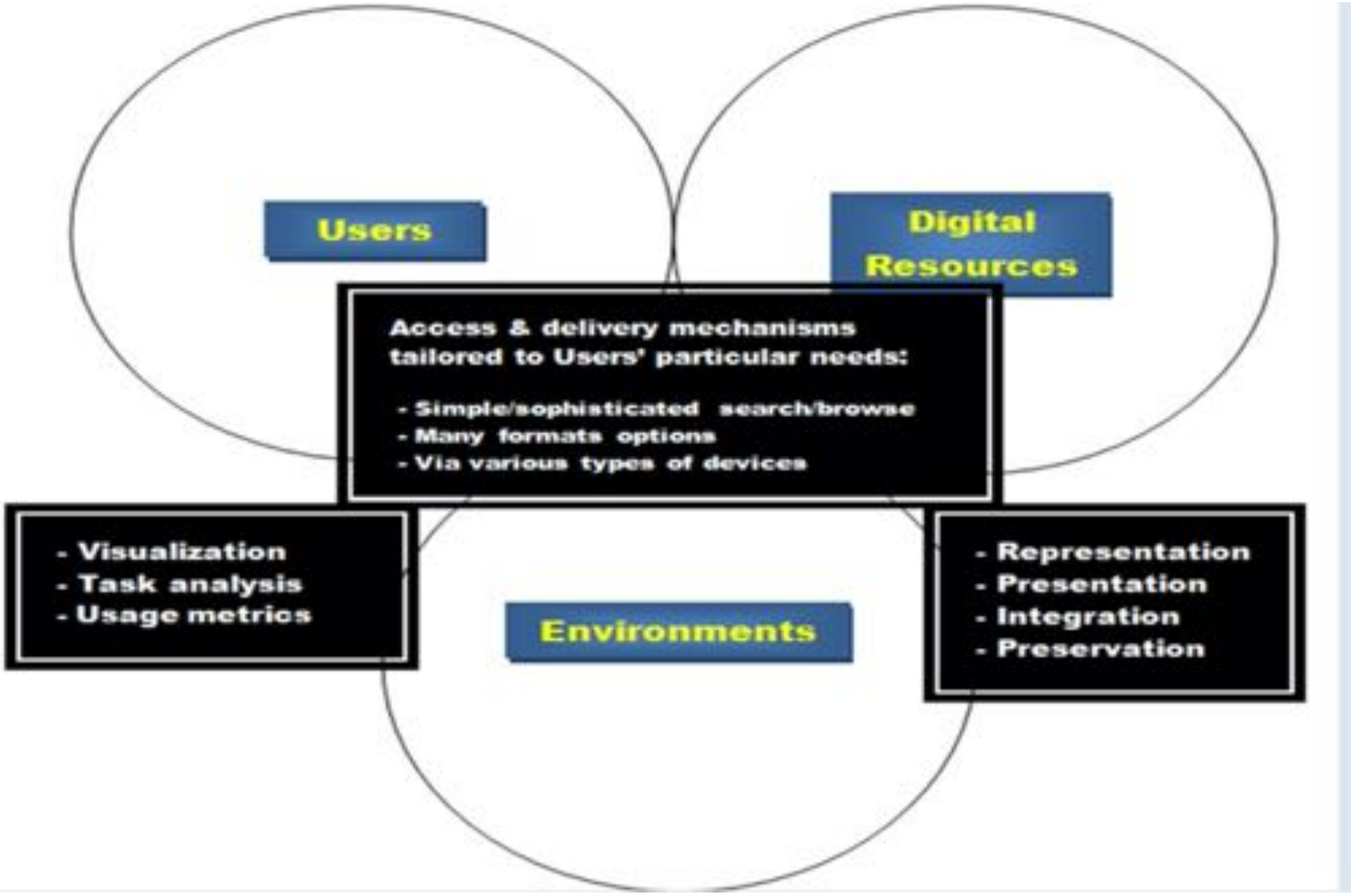
Logical Consistency/ Coherence

- *Are values in metadata elements consistent throughout?*
- *How does it compare with other data within the community?*

Digital Resources

- **A good digital collection is sustainable over time.**
- **Effective taxonomies add value and amplify the digital resources– allowing users to explore and delve deeper in multidimensional ways.**
- **Maintaining and ensuring quality requires a framework and cyclical process during the entire lifecycle.**
  - **-** interoperability, completeness, consistency, accuracy

Users, Digital Objects, and Environments

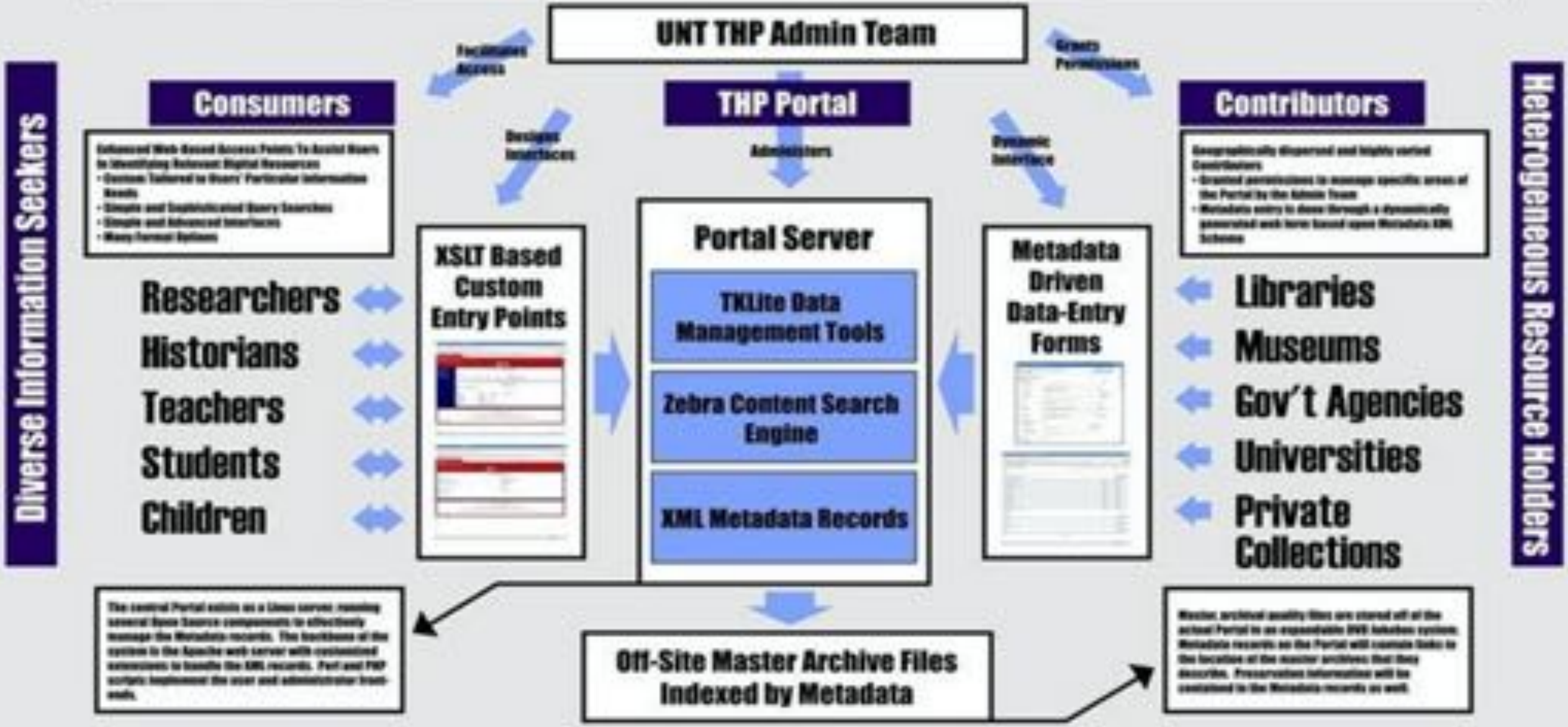


Digital Repositories

- Providing multiple formats access options
- Exposing the page level OCR text to an increasing number of search engines
- Allowing page turning interfaces and other interfaces designed for growing mobile devices use

The PORTAL *to* TEXAS HISTORY

Hosted by the University of North Texas Libraries

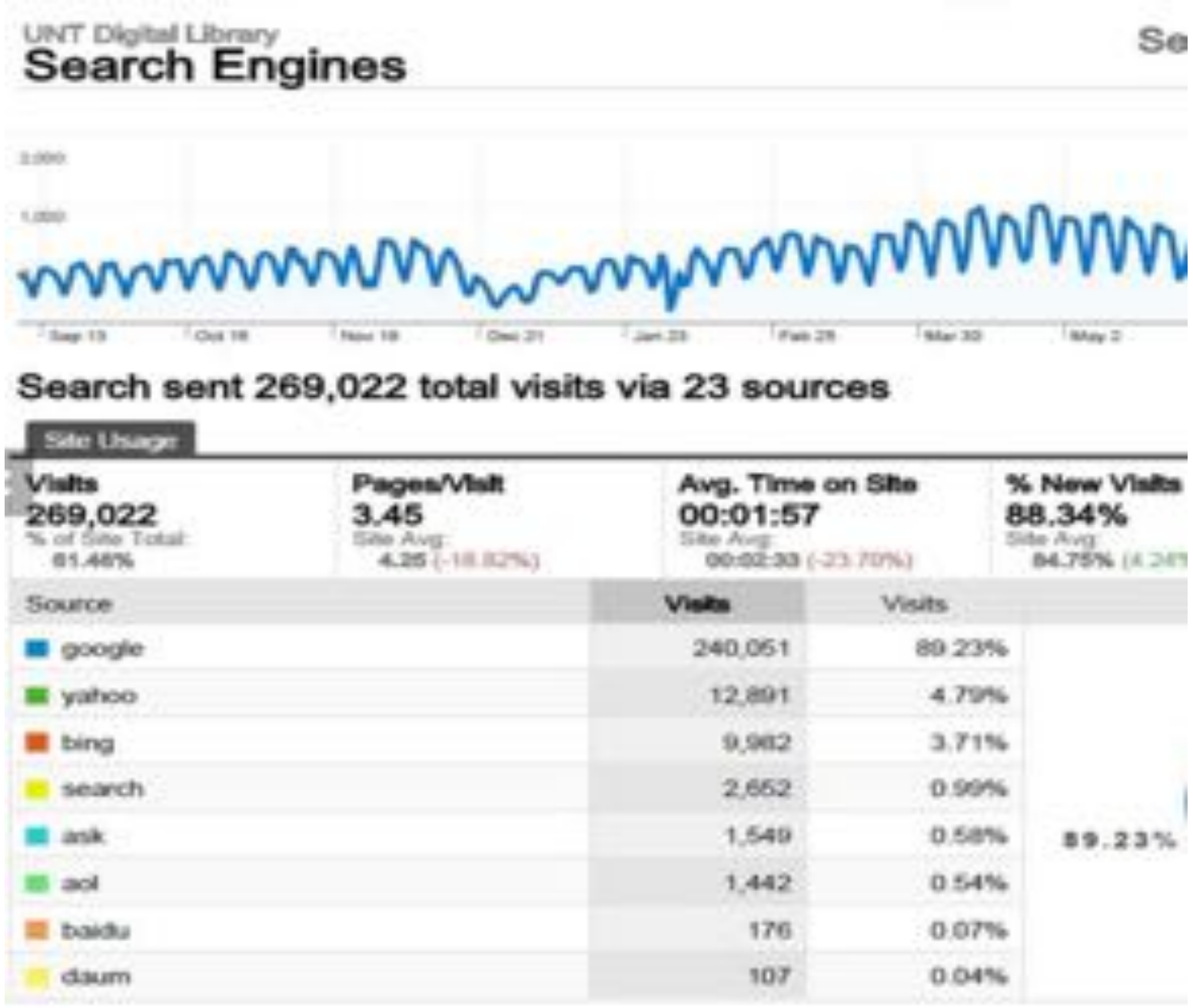


Tracking the Use of Digital Resources

- Tracking daily uses
- Countries coming from
- Types of devices used

Tracking the Use of Digital Resources

- Tracking referring sites
- Traffic sources
- Search engines used



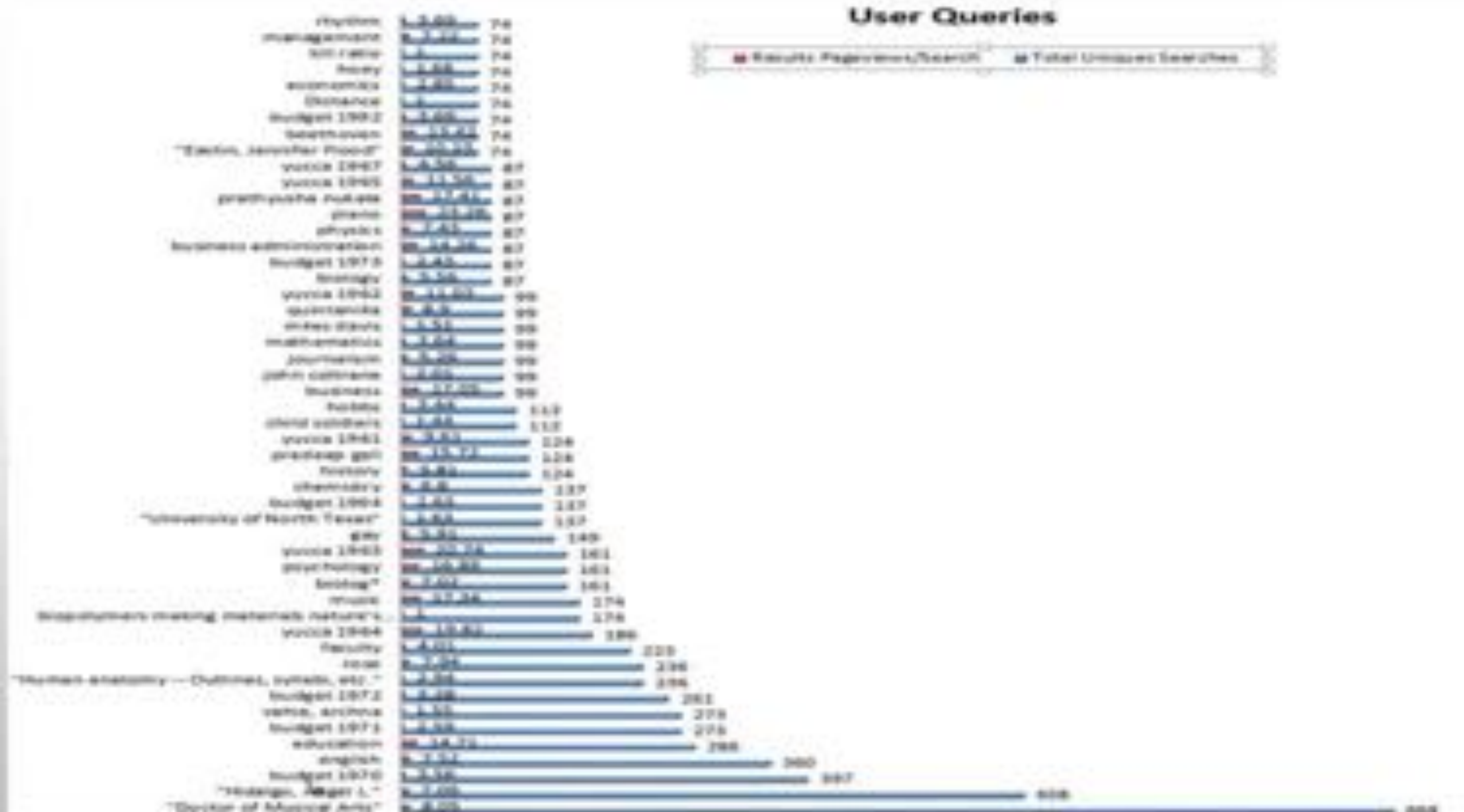
Top Browsers and Operating Systems Accessed one Collection

Browser		Users	New Users	Sessions	Bounce Rate	Pages / Session	Avg. Session Duration
Chrome		192,863	184,101	220,364	57.56%	2.52	0:02:08
Safari		64,930	62,742	73,431	63.97%	2.5	0:01:42
Firefox		41,733	39,474	45,728	58.92%	2.55	0:02:17
Internet Explorer		34,314	33,006	36,668	64.75%	2.13	0:01:41
Edge		13,911	13,507	15,920	58.70%	2.95	0:02:42
Opera Mini		6,183	6,135	6,718	69.38%	1.66	0:01:03
(not set)		4,887	4,887	2,687	98.25%	1	0:00:01
UC Browser		3,329	3,234	3,662	73.38%	1.53	0:00:58
Opera		2,402	2,318	2,949	60.09%	2.16	0:02:05
Android Browser		1,558	1,510	1,629	78.08%	1.34	0:01:10
Android Webview		1,367	1,355	1,546	68.43%	2.46	0:02:40
Safari (in-app)		1,284	1,248	1,391	60.68%	2.96	0:01:10
Samsung Internet		939	916	1,082	57.12%	2.22	0:01:46
Amazon Silk		713	701	761	71.88%	1.45	0:02:30

UNT's Case Study

- **To get a better sense of users' discovery of digital resources, we decided to assess and see:**
  - Whether users were arriving at our digital resources from searches that were answered by an items descriptive metadata or by parts of the full-text of the item.

- **This study analyzed access to UNTs ETD Collection from two sides:**
  - Searches that were answered by an items descriptive metadata
  - Users request met by parts of the full-text of the item.

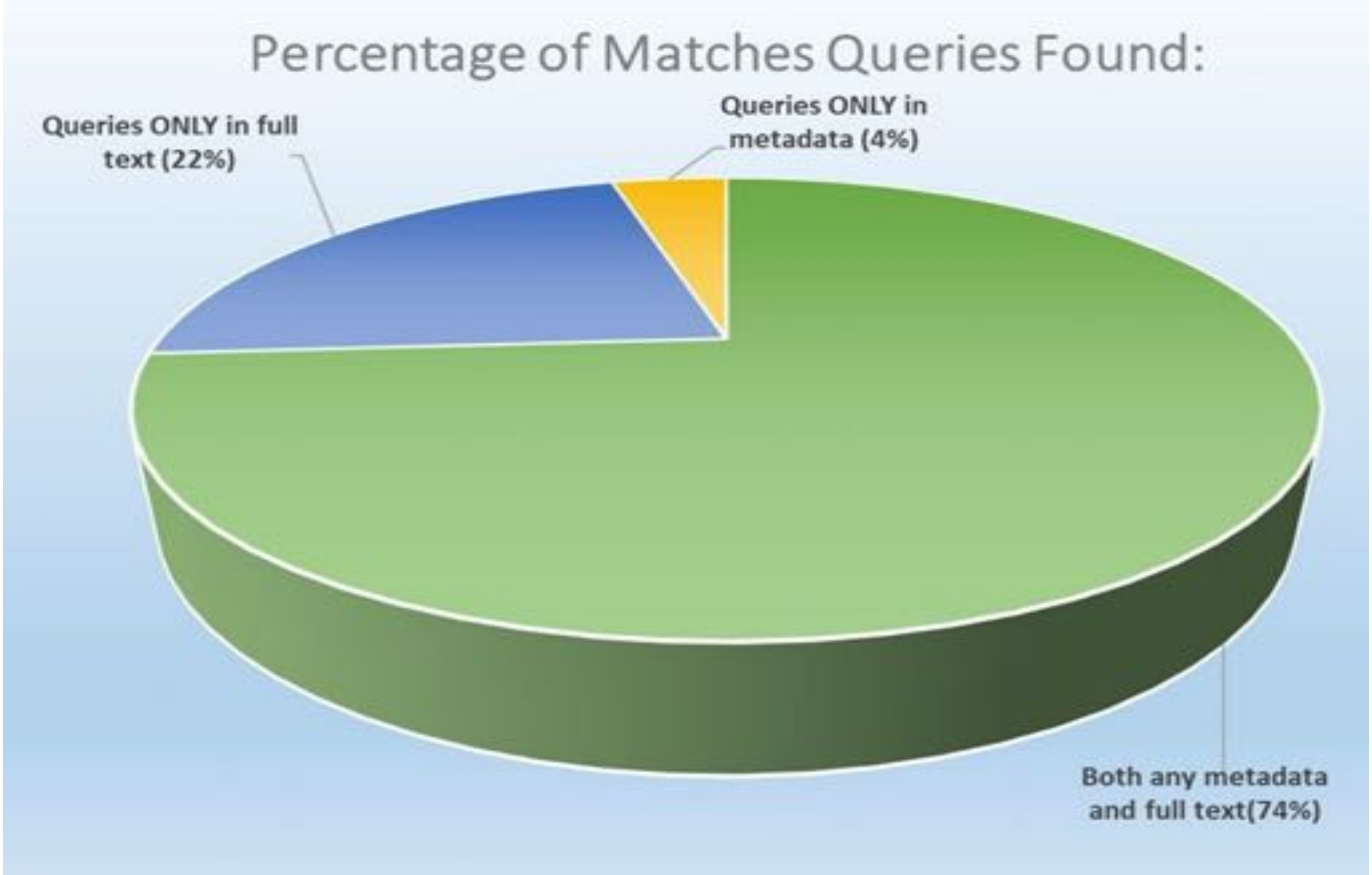

User Queries
Results Pageviews/Search
Total Unique Searches

Example Dataset Entries for Three Search Queries

Dataset Field	Example 1	Example 2	Example 3
Item	metadc129697	metadc146510	metadc155618
Query	susan cheal	human trafficking	article writing
Query Tokens	2	2	2
PageText	**0%**	**100%**	**100%**
Metadata	**100%**	**100%**	**50%**
Title	*0%*	*100%*	*50%*
Subject	*0%*	*100%*	*0%*
Agent	*100%*	*0%*	*0%*
Description	*0%*	*100%*	*0%*

Record Discoveries on Matches in Metadata and Full-Text (N=43420)

Matches found in: Bottom of Form	Total No. of Queries Found: %	%
Any part of query in full text	41,519	95.6%
Any part of query in metadata	33,779	77.8%
Both any metadata and full text	32,056	73.8%
100% of query in full text (all tokens)	36,318	83.6%
Queries ONLY in full text (but not in metadata)	9463	21.8%
100% of query in metadata (all tokens)	29661	68.3%
Queries ONLY in metadata (but not in full text)	1723	4.0%


Percentage of Matches Queries Found:
Queries ONLY in full text (22%)
Queries ONLY in metadata (4%)
Both any metadata and full text(74%)

A More Granular Level, Record Discoveries Broken Down by Match Percentages of Each Field

	0%	1-49%	50-74%	75-99%	100%	%>=1% found in field
Title	33,597	2,086	2,297	120	5,320	22.62%
Subj.	28,661	1,591	1,736	61	11,371	33.99%
Agent	29,276	193	293	4	13,654	32.57%
Descr.	29,274	3,048	3,454	350	7,294	32.57%

This Table shows record discoveries per field or the extent of the matches (partially for longer query strings [up to 31 tokens]) and the overlap across multiple fields (N=43420)

Trends

- **Technological innovation**
  - Web 3.0 (*built using artificial intelligence, machine learning and the semantic web*)
  - Big Data as Information Assets (*The #Vs of Big Data*)
  - Evolving formats and proliferation of non-text items (*Flickr, Youtube, Netflix,…)*
  - Discovery tools (*enhanced access capability*)

- **The Semantic Web**
  - The Resource Description Framework (RDF) provides the basic capabilities to define knowledge-based objects on the Internet with basic features such as Is-A relations and object properties.
  - The Web Ontology Language (OWL) adds additional semantics and integrates with automatic classification reasoners.

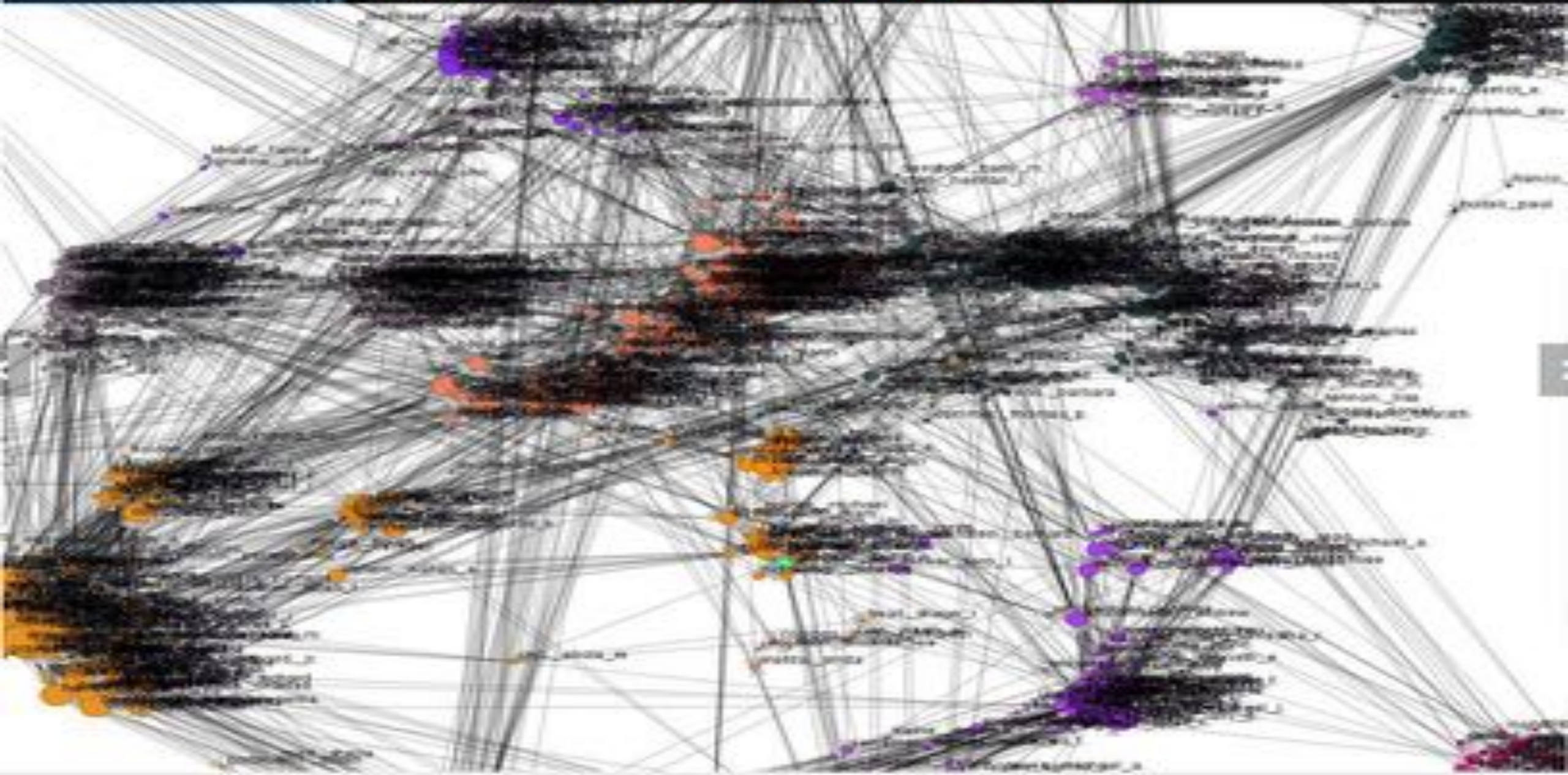
Trends…

- **User Centered**
  - Evolving user needs and requirements
  - Folksonomies and related components of collaborative information services
  - Leverage and take advantage the available tools and best practices.

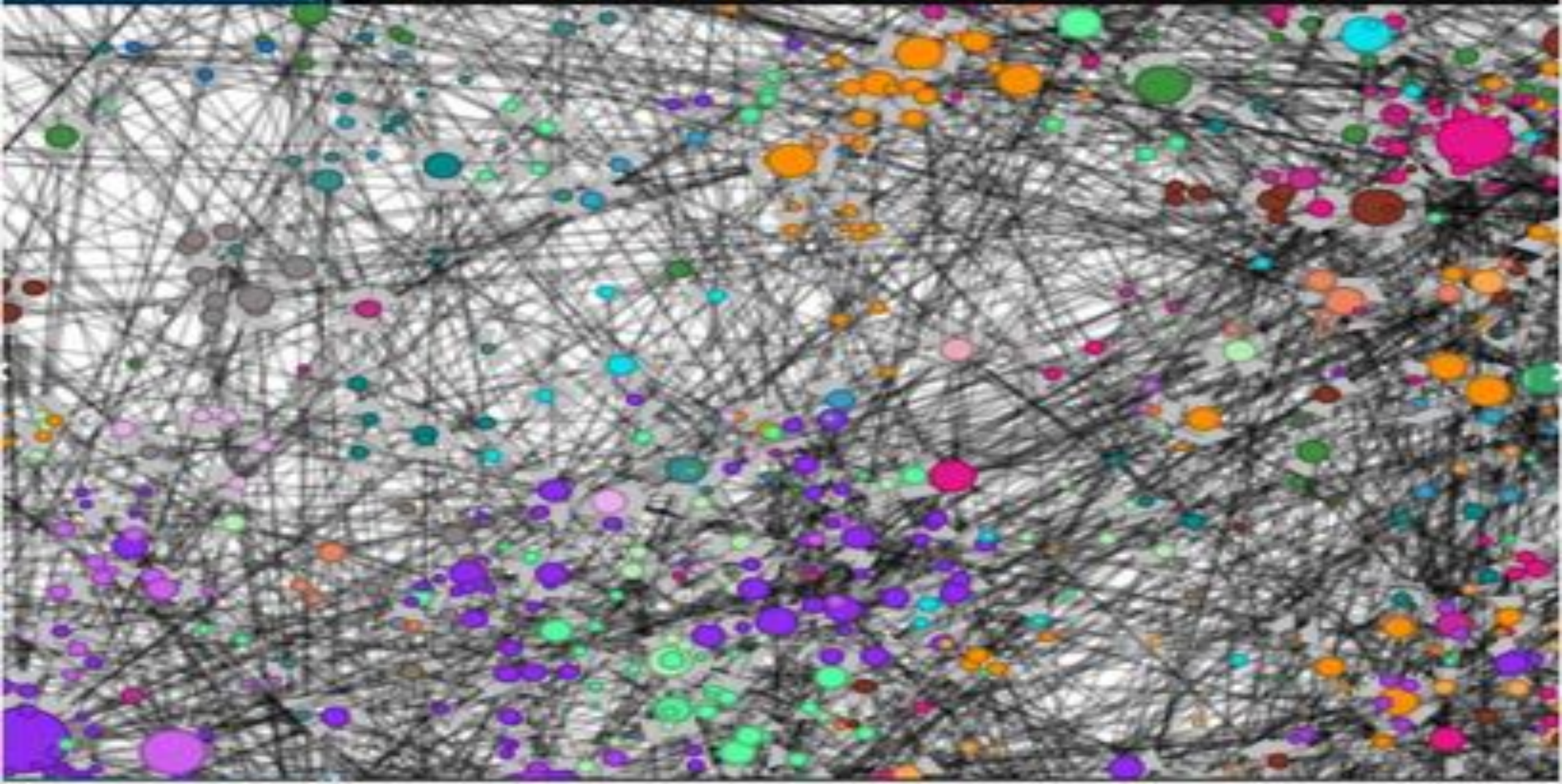
- **Skills and training**
  - Competencies and marketable skills
  - Diverse potential employers
  - Growing number of non-traditional jobs.


ALSO
BIG
DATA
TOOLS
DISK
TARGET
APPLIED
SENSOR
MAY
EXAMPLES
SHARED
DEFINITION
MOVING
EVERY
WITHIN
PRACTITIONERS
CAPTURE
BUSINESS
INTERNET
SETS
GENOMICS
DISTRIBUTED
MANAGEMENT
TERABYTES
ABILITY
PETABYTES
INCLUDE
SYSTEMS
TOLERABLE
SIZE
STORAGE
PARALLEL
SAN
MASSIVELY
GROW
SET
RELATIONAL
LARGER
CONTINUES
SOCIAL
TENS
COMPLEX
CITATION
ANALYTICS
BIOLOGICAL
PROCESSING
CAPACITY
HUNDREDS
NETWORKS
DATABASES
RECORDS
SEARCH
GARTNER
WORKING
DIFFICULTY

Data Visualization



Subject Terms Visualization



Summary

- A good digital collection is sustainable over time, which is to say its individual items are curated and actively managed during their entire lifecycle in both a trusted and cost-effective manner.
- With proper indexes and metadata, successful digital curation will mitigate digital obsolescence, keeping the information accessible to users indefinitely.
- Considering the multiple stakeholders in the digital ecosystems, a collaborative approach is the best way to addressing information organization challenges in general!

Summary

- The shift from a traditional library science focus to the broader information science focus required IS programs to prepare students for jobs outside the traditional market and possibly for jobs that might not exist today.

- Professional organizations like ASAIB have provided forums in which information scholars, researchers, educators, professionals, and publishers could share their insights on the ever-evolving horizon in the field of indexing and of course, library and information sciences at large.

Thank you!

Questions?

Association of Southern African Indexers and Bibliographers (ASAIB)

Excellence from A to Z

Daniel.Alemneh@ac.up.za
Daniel.Alemneh@unt.edu

Cited Works

Alemneh, D. and Phillips, M. (2018). Metadata versus Full-Text: Tracking Users' Electronic Theses and Dissertations (ETDs) Seeking Behavior. iConference-2018, Sheffield, UK, March 25-28, 2018. Retrieved from: https://digital.library.unt.edu/ark:/67531/metadc1132756/

Alemneh, D. and Phillips, M. (2016). Indexing Quality and Effectiveness: An Exploratory Analysis of Electronic Theses and Dissertations Representation. The Association for Information Science and Technology (ASIS&T) Annual Conference, October 27, 2016; Silver Springs, MD. Retrieved from: https://digital.library.unt.edu/ark:/67531/metadc957455/

Alemneh, D. and Rorissa, A. (2014). Facilitating Discovery and Use of Digital Cultural Heritage Resources with Folksonomies: A Review. in Annual Review of Cultural Heritage Informatics 2012-13. PP. 17-29, Facet Publishing. Retrieved from: https://digital.library.unt.edu/ark:/67531/metadc1128928/

Bruce, T.R., & Hillmann, D.I. (2004). The continuum of metadata quality: defining, expressing, exploiting. In Hillman, D. and Westbrook, L. (Eds.), *Metadata in Practice.* Chicago: American Library Association, pp. 238-256.

Digital POWRR) (2016). *Digital POWRR Webinar Series*. Retrieved from: https://digitalpowrr.niu.edu/institutes/survived-powrr-wkshp/

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 3*(1), 134-140

King, S. et al. (2018) Revisiting Indexing & Abstracting in the Digital Era. *MIRS Symposium*, University of North Texas, February 12, 2018. Retrieved from: https://digital.library.unt.edu/ark:/67531/metadc1164546/

Klein, M., & Noy, N. (2003). A component-based framework for ontology evolution. In *Workshop on Ontologies and Distributed Systems*. Retrieved from http://www.researchgate.net/publication/2930642_A_Component-Based_Framework_for_Ontology_Evolution

Moen, W.E., Stewart, E.L, & McClure, C.R. (1998). *The Role of Content Analysis in Evaluating Metadata for the U.S. Government Information Locator Service (GILS): Results from an Exploratory Study*. Retrieved from: http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm.

Netcraft (2022*). Netcraft Web Server Survey. Retrieved from*: https://news.netcraft.com/archives/category/web-server-survey/

Névéol, A., Shooshan, S. E., Humphrey, S. M., Mork, J. G., & Aronson, A. R. (2009). A recent advance in the automatic indexing of the biomedical literature. *Journal of Biomedical Informatics*, 42(5), 814-823. doi:10.1016/j.jbi.2008.12.007

NISO (2021) *ANSI/NISO Z39.4-2021 Criteria for Indexes*. Retrieved from: https://www.niso.org/publications/z394-2021-indexes

NISO (2019). *Bridging the Gap Between Abstracting & Indexing Provider Needs and Discovery Service Approaches*. Retrieved from:

https://www.niso.org/publications/odi-bridging-gap

Park, J. & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly*, *48* (8), 96-715.

Rorissa, A. (2007). Benchmarking Visual Information Indexing and Retrieval.. *ASIS&T Bulletin,* February/March 2007. Retrieved from:

https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/bult.2007.BULT1720330310

Thornburg, G., & Oskins, M. (2007). Misinformation and bias in metadata processing: Matching in large databases. *Information Technology and Libraries, 26* (2), 15-26.

Zavalina, O. et al. (2015). Building a Framework of Metadata Change to Support Knowledge Management. International Knowledge Management Conference (ICKM-2015), April 17, 2015; [Singapore]. Retrieved from: https://digital.library.unt.edu/ark:/67531/metadc505014/m1/