

# Enhancing ICPSR's Subject Thesaurus for the Artificial Intelligence Era

Megan Chenoweth, Senior Associate Librarian, ICPSR  
Jared Lyle, Director of Metadata and Preservation, ICPSR

# Thesauri and the ICPSR Subject Thesaurus

- Background on thesauri
  - Controlled vocabulary
  - Relationships between terms
- The ICPSR Subject Thesaurus
  - Terms related to social science and health
  - Used for indexing ICPSR data collections
  - Users can browse, search, and filter by subject terms

## implants

Search for this term

**Scope Notes:** Devices surgically implanted into the body; includes artificial joints, breast implants, dental implants, pacemakers, etc.

### Broader Terms

Term

[assistive devices](#)

### Narrower Terms

No narrower terms found

### Related Terms

Term

[biomaterials](#)

[surgery](#)

### Preferred Term

No preferred terms found

### Non-Preferred Term

Term

[medical implants](#)

<https://www.icpsr.umich.edu/web/ICPSR/thesaurus/10001/terms/25782>

# Challenges and Questions

- Thesaurus Quality
  - Updated only upon request and not comprehensive
  - Newly added terms not added retroactively
  - Non-thesaurus terms appear in studies
- Relevance
  - Discovery via large language models is increasingly common. Do thesauri still help?



# **Subject Thesaurus As-Is Assessment and Initial Findings**

# What Does the Thesaurus Look Like?

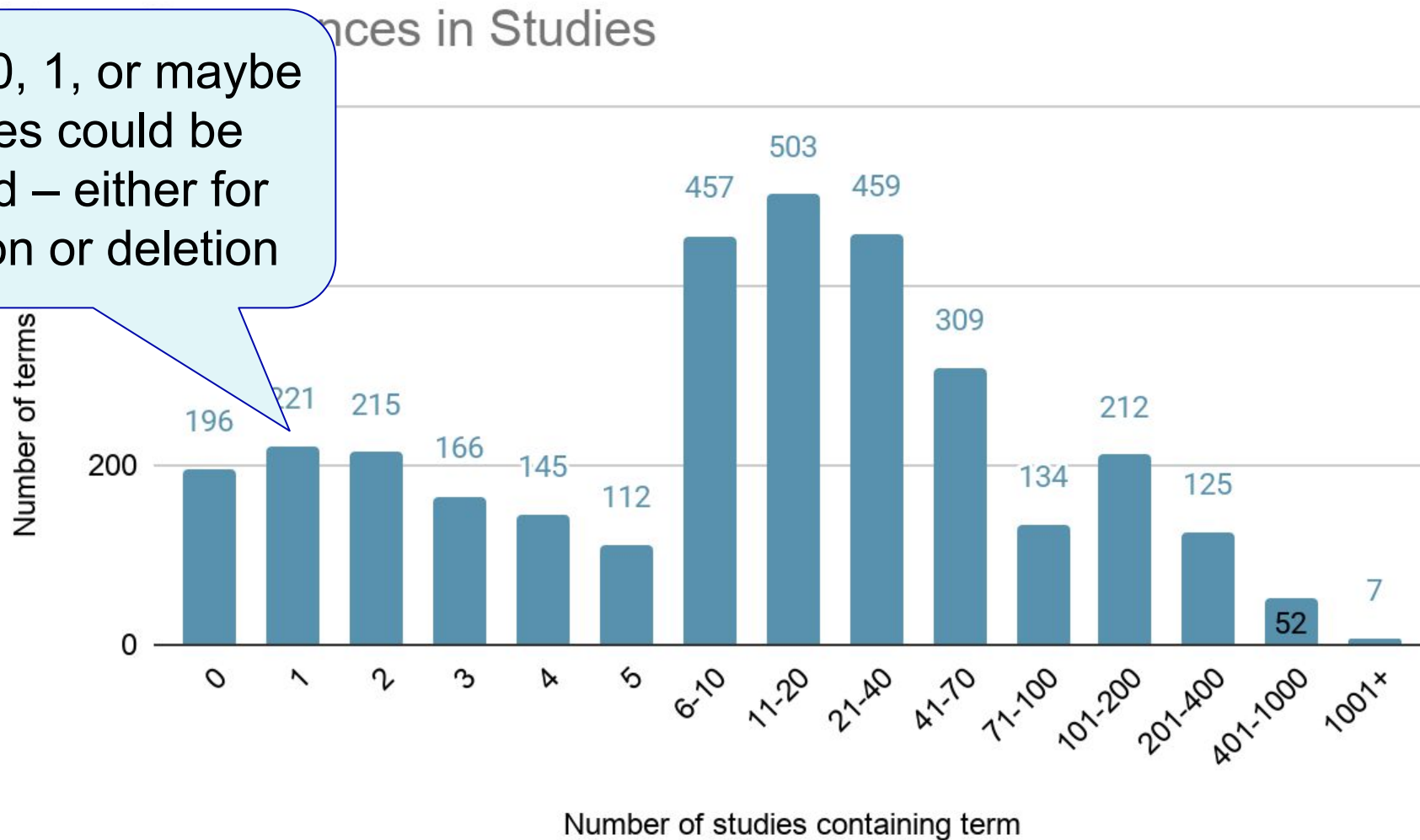
Type of Term	n	%
Preferred terms	3313	87.3%
Terms with relationships		
Has broader/narrower term	2009	52.9%
Has related terms	2877	75.8%
No broader/narrower or related terms	66	1.7%
Terms with scope notes and descriptors		
Has a scope note/descriptor	752	22.7%
No scope note/descriptor	2561	77.3%
Non-preferred terms	482	12.7%
<b>Total</b>	<b>3795</b>	<b>100%</b>

Adding relationships for these terms might improve their findability for users

Defining these terms would support consistent application and discovery

# How Are Terms Distributed Across Studies?

Terms in 0, 1, or maybe 2 studies could be reviewed – either for expansion or deletion



# Do Study Subject Terms Match Thesaurus Terms?

Match Status	Curated Studies	Self-Published Studies
Thesaurus term	150,414 (96.1%)	8,871 (25.7%)
Non-preferred term	410 (0.3%)	1,030 (3.0%)
Term not in thesaurus	5,725 (3.7%)	24,602 (71.3%)
<b>Total</b>	<b>156,549 (100%)</b>	<b>34,503 (100%)</b>

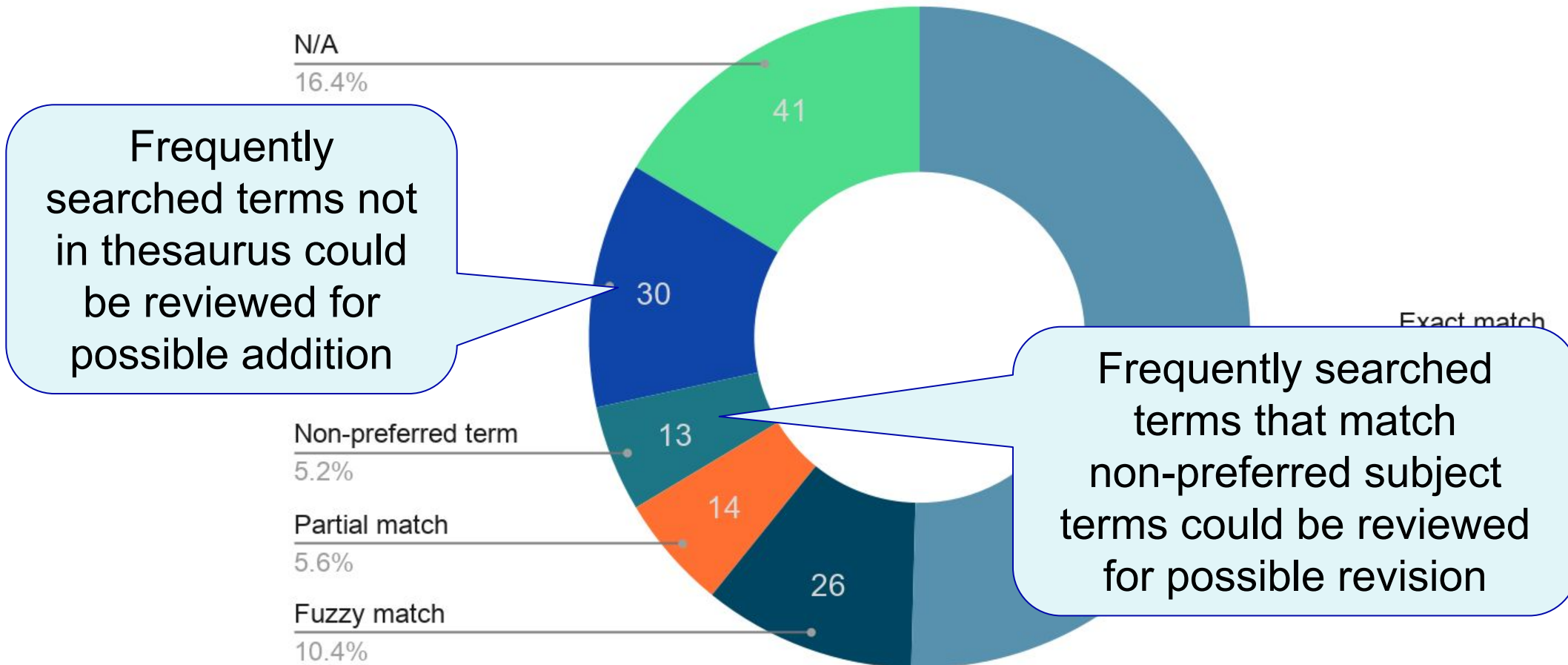
Non-preferred terms  
can be replaced with  
preferred terms

Non-matching terms  
can be evaluated and  
either added to the  
thesaurus or replaced

Terms not in thesaurus  
can be disallowed for  
self-publishing in the  
future

# Do Subject Terms Match What Users Search For?

Occurrences of 250 Most Frequent Search Terms in Thesaurus

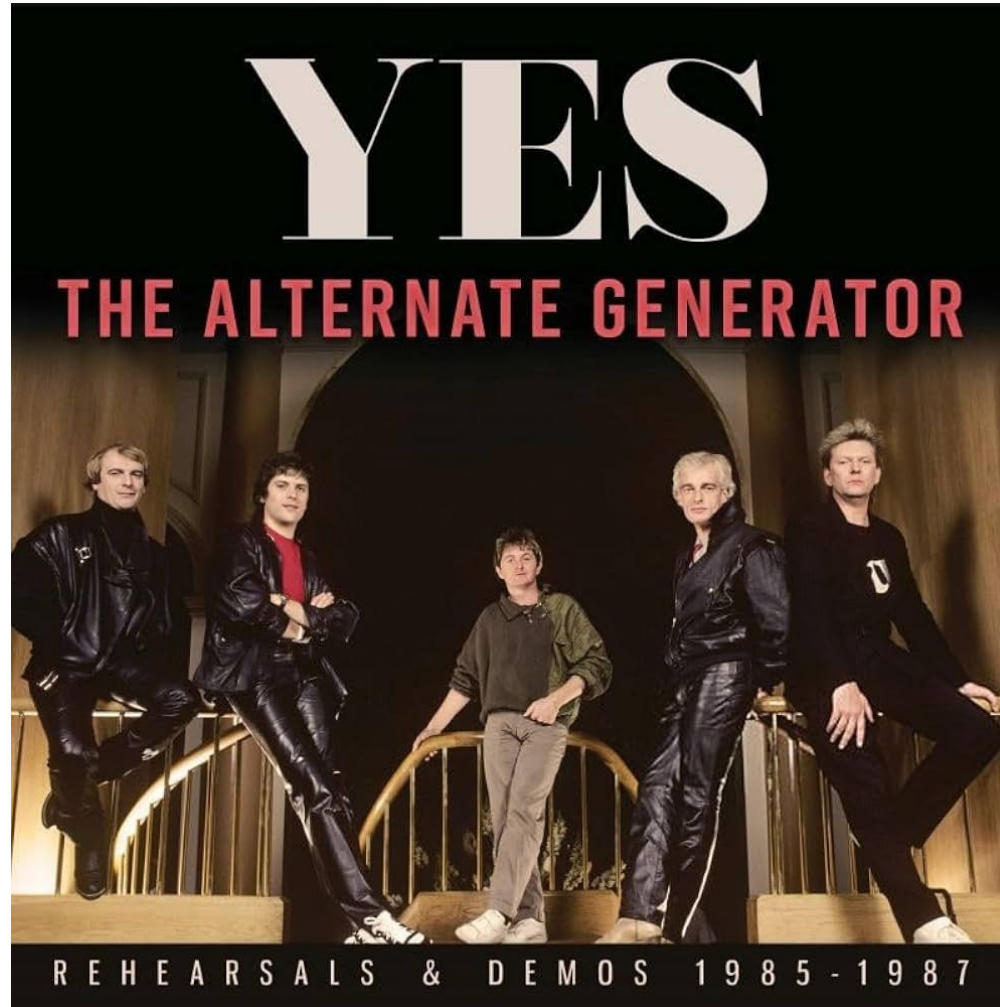






# **It's the AI Age. Do Subject Thesauri Still Matter?**

# Do Subject Thesauri Still Matter?



<https://www.amazon.com/Alternate-Generator-Yes/dp/B0D5J2XDYV>

# Generative AI Needs High-Quality Metadata

- Training data are messy. Consistent terminology makes for better trained models.
- AI is probabilistic, so generative AI can hallucinate. Clear categories and definitions can help.
- Models need context. Thesauri can provide it.

# Subject Thesauri and AI: Applications

- We can't control how commercial LLMs use subject thesauri
- We can influence their use in local LLM instances
  - Retrieval augmented generation (RAG)
  - Query expansion
  - Example: BMQExpander (<https://arxiv.org/pdf/2508.11784>)
- Organizations developing local LLMs should experiment with how best to use knowledge embedded in thesauri and controlled vocabularies

# Conclusion and Next Steps

- The thesaurus isn't a dinosaur
- So let's keep it modern
  - Add terms based on user searches
  - Expand or weed terms used 0, 1, or 2 times
  - Enhance terms with context: scope notes, synonyms, acronyms, and provenance
- Partner with teams implementing local LLMs for discovery to see how the thesaurus can help



Credit: [FossilFuel on Etsy](#)

# Learn More



**WEBSITE**

[www.icpsr.umich.edu](http://www.icpsr.umich.edu)



**EMAIL**

[mmchenow@umich.edu](mailto:mmchenow@umich.edu)