MICHEL BIEZUNSKI, INFOLOOM

INTEGRATING INDEXES WITH TOPIC MAPS

A PROJECT WITH THE NYU LIBRARY

American Society for Indexing Conference

April 28, 2018 Cleveland, OH

WHO I AM

- Michel Biezunski is an innovator and developer who cares about human knowledge. Background is in science and publishing.
- Co-authored the Topic Maps standard (ISO/IEC 13250).
- Infoloom helps its clients leverage and add value to their information by providing tools and consulting services.

CONTENTS

- Indexes and Topic Maps
- Integrating Indexes for the NYU Library.
- History, Design and Challenges
- Topic Curation Tasks

TOPIC MAPS STARTED WITH INDEXES

- Around 1991, the Open Software Foundation was trying to integrate the documentation of its members' products: Motif, DCE, OSF/1, and others. They decided to go for "interoperable indexes".

- At the same time, a complex hyperlinking-based standard (HyTime ISO/IEC 10744) was just released, and a study group started to look into how to use it.

FORK

- This working group eventually split into 2 groups :

1. The Docbook Document Type Definition

2. Topic Maps

THE DOCBOOK APPROACH

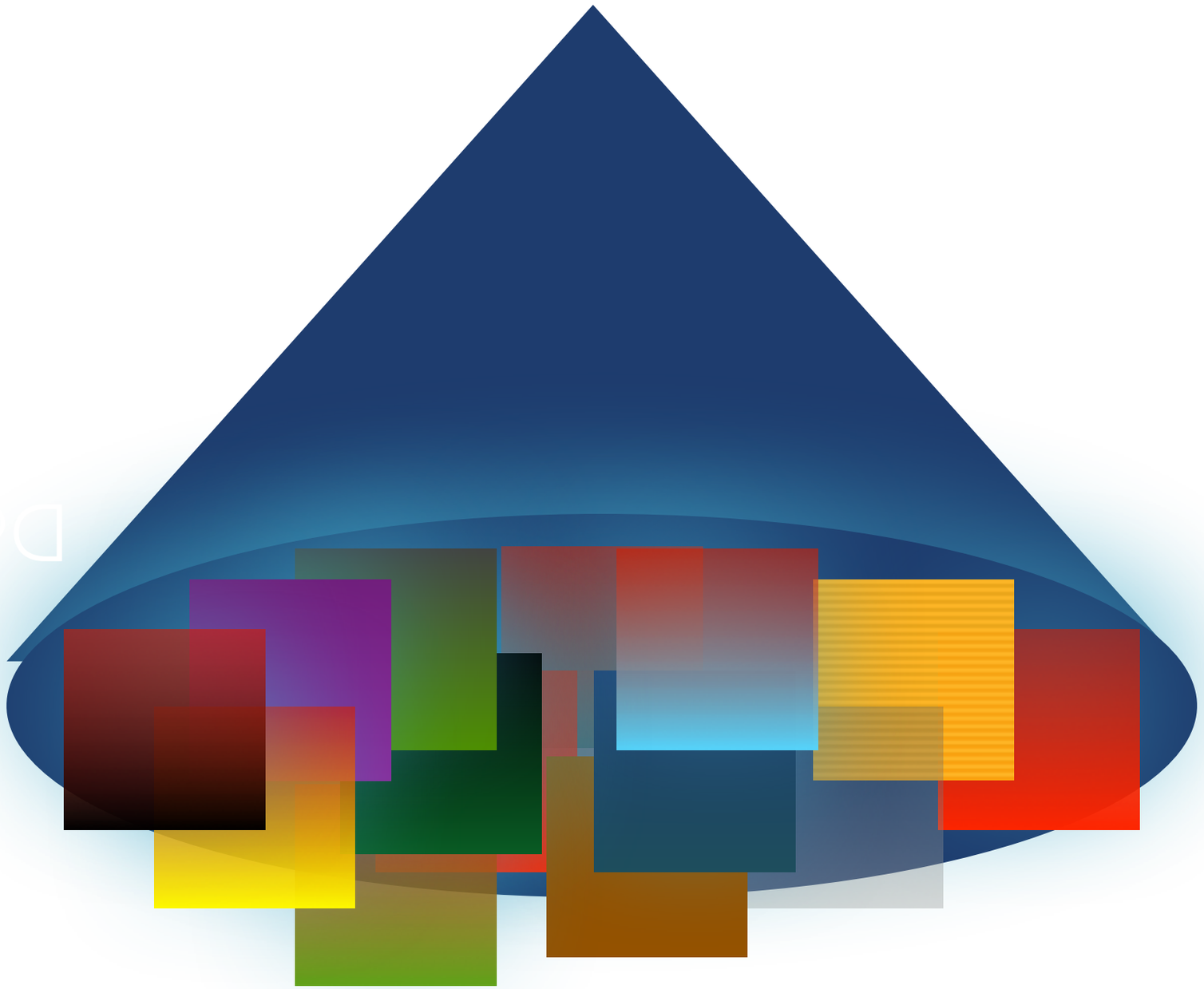
- Indexes are described with generic markup within the document.

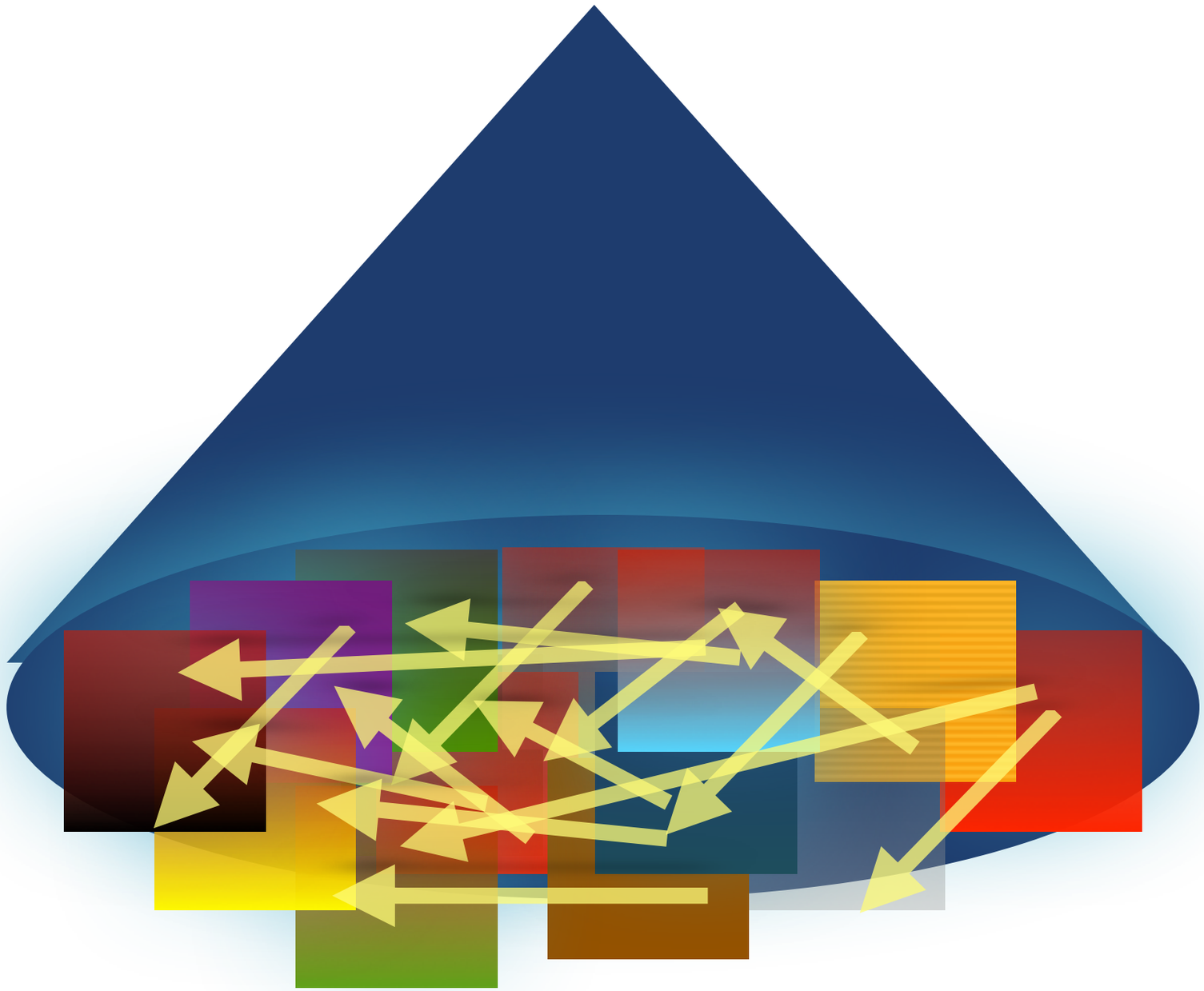
```
<indexterm>
  <primary>information</primary>
  <seealso>data</seealso>
  <secondary>dissemination</secondary>
</indexterm>
```

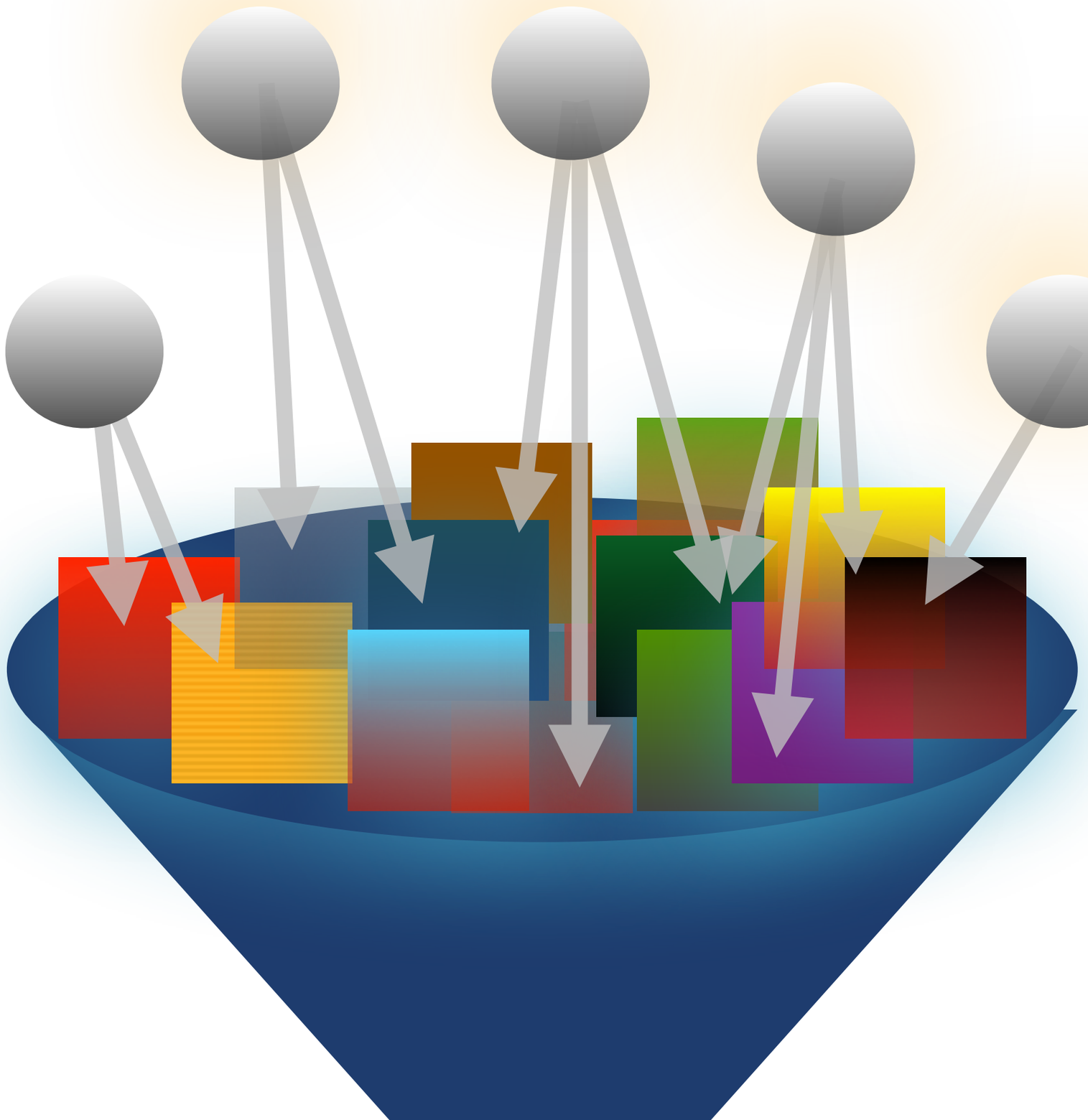
- The index is automatically created by Docbook tools.

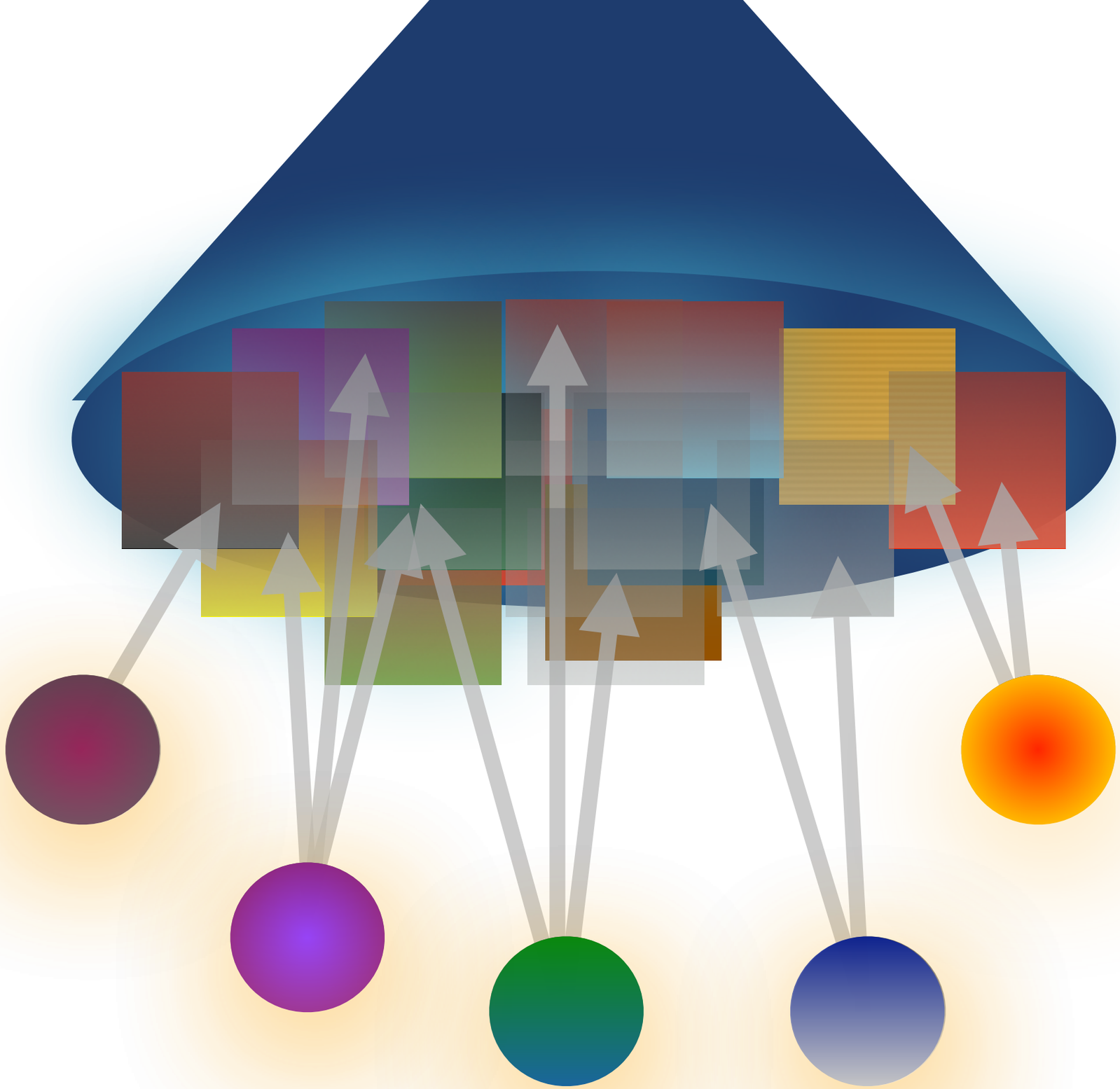
THE TOPIC MAPS APPROACH

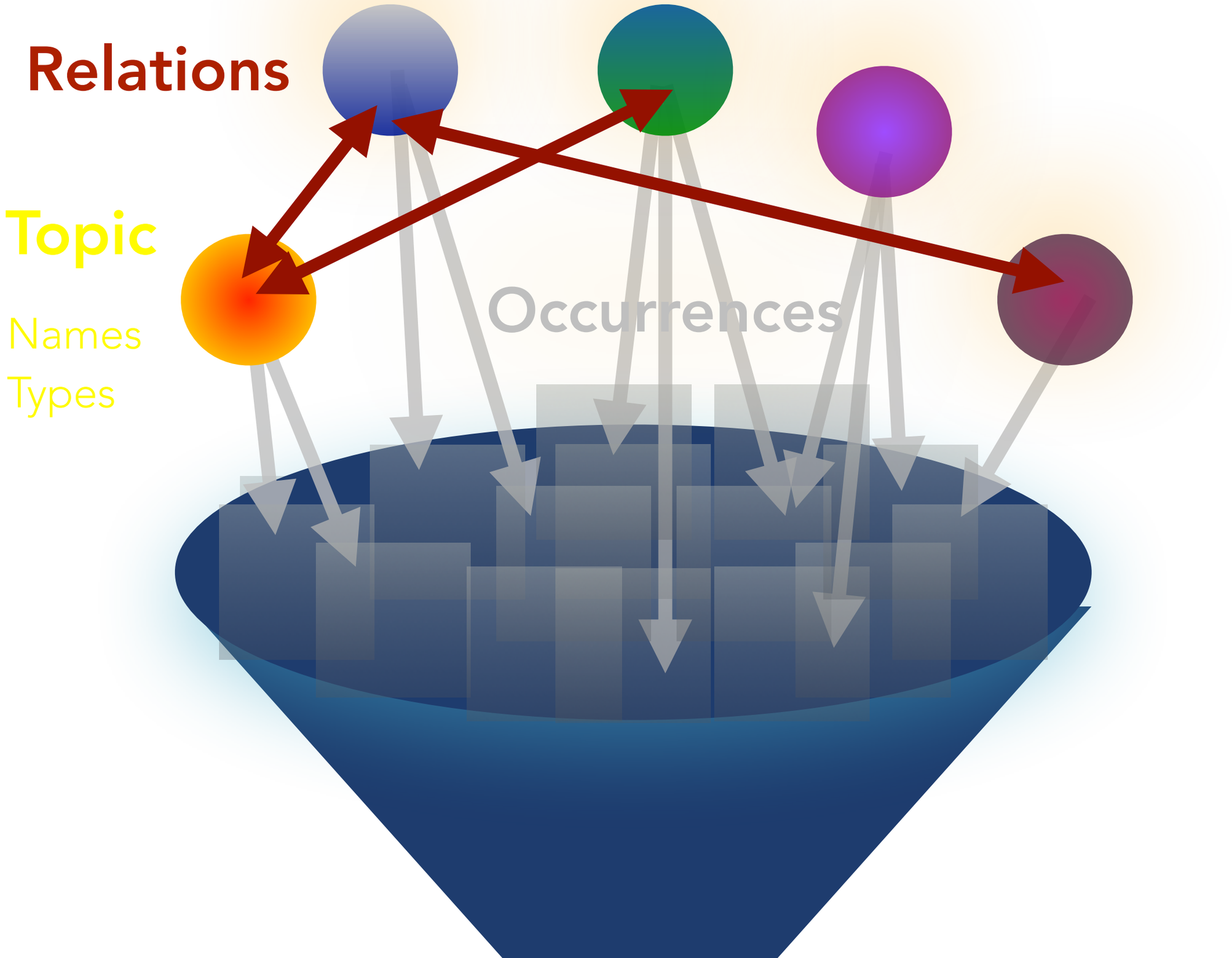
- Philosophically:
  - Topics are abstract units of meaning.
  - The semantic space is a map, where topics are nodes.
- Technically :
  - A topic is a computer-representable proxy for a subject.
  - Topics are objects with properties.

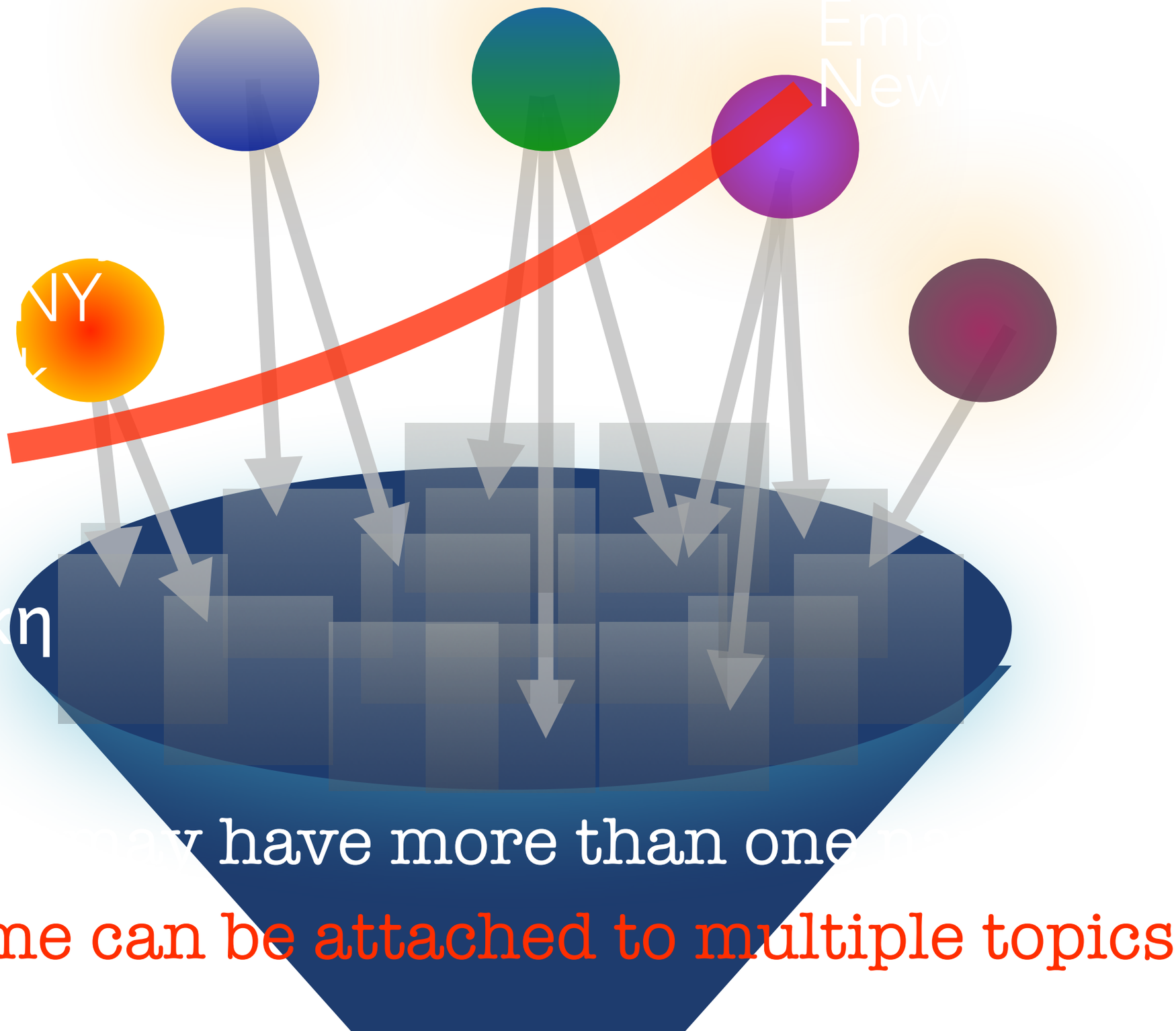









Relations
Topic
Names
Types
Occurrences



may have more than one name

One name can be attached to multiple topics.

TOPIC MAPS ARE NOT NEW

- **Indexes** : List of topic names with occurrence indicators (in name alphabetic order)
- **Tables of Contents**: List of topics (headers), followed by their occurrences (in document order)
- **Cross-references**: Link between two (or more) occurrences of the same topic.
- **Glossaries, Dictionaries** : List of topic names and their occurrence content serving as definition.
- **Thesaurii**: Relations between topics
- **Library Catalogs, Web Taxonomies, Ontologies, Controlled Vocabularies,** etc.

Each of these finding aids is a pre-defined query in a topic map database.

THE GENESIS OF THE NYU ENM PROJECT

A book index

A library catalog

is a topic map, with human-created topics

"Enhanced Networked Monograph Project", funded by Carnegie Mellon: Integrate indexes of ~100 ebooks and navigate through the result.

THREE MAIN STEPS

1. Extract topics from the indices

2. Build a curation tool to manage the topic map.

3. Provide data for a navigation system.

THREE CHALLENGES

- Extraction challenges
- Semantic Integration challenges
- Reproductibility challenges

EXTRACTING TOPICS FROM EBOOKS

```
<h2 class="h2a"><a id="page_201"></a><a id="index"></a>INDEX</h2>
<p class="indexmain1">affirmative action, <a class="nounder"
href="ch05.html#page_107">107</a>&#8211;108, <a class="nounder"
href="conclusion.html#page_155">155</a></p>
<p class="indexmain">American Civil Liberties Union, <a class="nounder"
href="ch02.html#page_47">47</a></p>
<p class="indexmain">Amherst, <a class="nounder" href="ch05.html#page_109">109</a></p>
<p class="indexmain">antideterminism, <a class="nounder" href="ch05.html#page_89">89</a>, <a
class="nounder" href="ch05.html#page_97">97</a>, <a class="nounder"
href="ch05.html#page_101">101</a></p>
<p class="indexmain">apartheid, <a class="nounder" href="ch02.html#page_56">56</a></p>
<p class="indexmain"><em>Armour v. Salisbury</em>, <a class="nounder"
href="introduction.html#page_10">10</a></p>
<p class="indexmain">arrest rates. <em>See</em> crime rates</p>
<p class="indexmain">assumed risk, <a class="nounder" href="ch05.html#page_112">112</a></p>
<p class="indexmain">aversive racism model, <a class="nounder"
href="ch06.html#page_128">128</a>&#8211;129</p>
```

case of entry

range

```
<p class="indexmain">AMA (American Medical Association), <a class="nounder"
href="intro.html#page_3">3</a>, <a class="nounder" href="ch07.html#page_181">181</a>&#8211;82, <a
class="nounder" href="ch07.html#page_183">183</a></p>
<p class="indexsub">Black membership, denial of, <a class="nounder"
href="ch01.html#page_38">38</a></p>
<p class="indexsub">and hospital desegregation, <a class="nounder"
href="ch01.html#page_41">41</a>&#8211;42</p>
<p class="indexsub">Medicare, opposition to, <a class="nounder" href="intro.html#page_19">19</a></p>
<p class="indexsub">national health insurance, opposition to, <a class="nounder"
href="ch07.html#page_181">181</a></p>
<p class="indexsub"><a id="page_324"></a>and population control, <a class="nounder"
href="ch06.html#page_156">156</a>&#8211;57</p>
<p class="indexsub">and segregation, <a class="nounder" href="ch07.html#page_182">182</a></p>
<p class="indexmain">Anaheim, CA, <a class="nounder" href="ch03.html#page_83">83</a></p>
```

subentries

EXTRACTING DESIGN DECISIONS

- Entries and subentries
- Page numbers and ranges
- See and See also relations
- Lowercase/Uppercase entries
- Lexical Recognizers
- Automatic Relations

EXTRACTION CHALLENGES

- Parsing errors with page numbers and entries/ subentries/see/seealso delimiters
- Epub indexes were not structured enough.
- Each book used a slightly different pattern, that needed to be analyzed before extracting.

EPUB3 INDEX STRUCTURE

http://www.idpf.org/epub/idx/

entry
- term
- locator

xref-preferred
xref-related

entry
- term
- entry-list
- entry
- …

SEMANTIC INTEGRATION CHALLENGES

- Several terms for the same topic.
  Ex: Bill Clinton vs. William Jefferson Clinton

- One term used for several topics.
  Ex: “New York” for “New York City”, “New York State”.

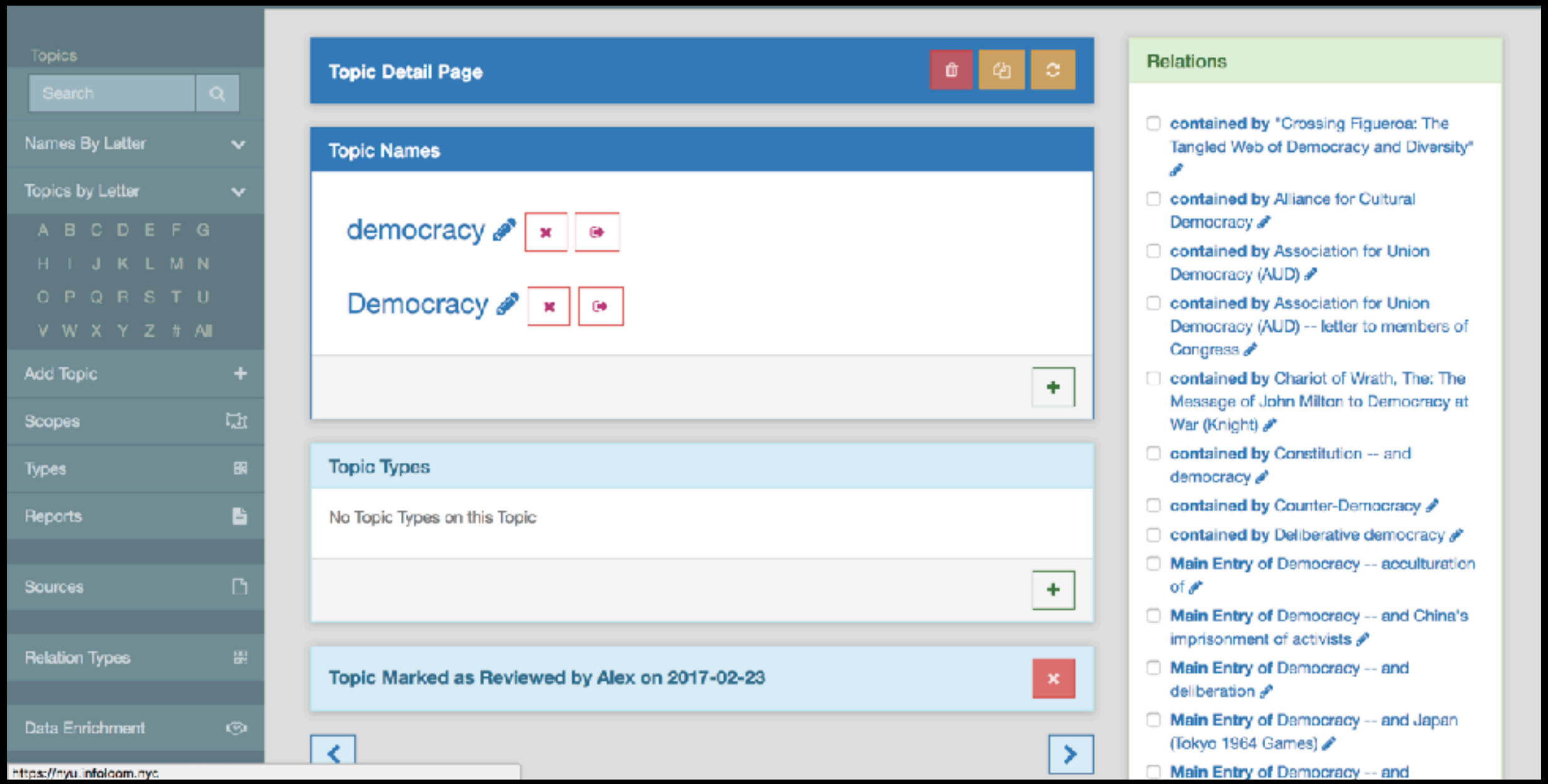
- Level of topic granularity depends on each book
  Ex: “Racism” may be missing as an index entry if the book is called “Racism in America”

- However: Mess is a feature, not a bug.

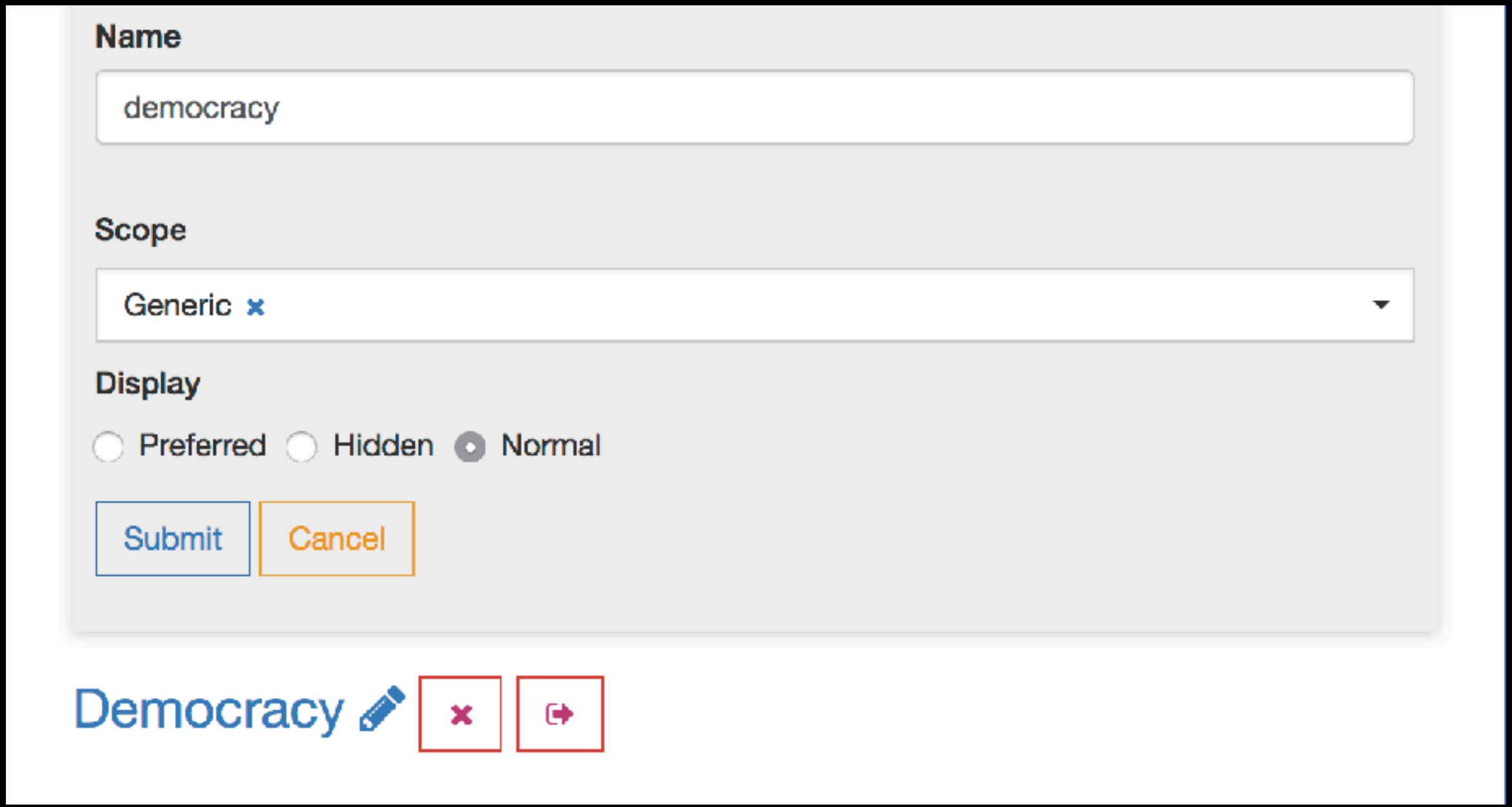
TOPIC CURATION TOOLKIT

- Creates, delete, edit
    - topics
    - topic names
        - Disambiguation using “scopes”
        - Languages
    - relations between topics
    - occurrences with locators
    - occurrence types
    - links to external web sites
    - links to controlled vocabularies

TOPIC PAGE


Topics
Search
Names By Letter
Topics by Letter
A B C D E F G
H I J K L M N
O P Q R S T U
V W X Y Z # All
Add Topic
Scopes
Types
Reports
Sources
Relation Types
Data Enrichment
Topic Detail Page
Topic Names
democracy
Democracy
Topic Types
No Topic Types on this Topic
Topic Marked as Reviewed by Alex on 2017-02-23
Relations
contained by "Crossing Figueroa: The Tangled Web of Democracy and Diversity"
contained by Alliance for Cultural Democracy
contained by Association for Union Democracy (AUD)
contained by Association for Union Democracy (AUD) -- letter to members of Congress
contained by Chariot of Wrath, The: The Message of John Milton to Democracy at War (Knight)
contained by Constitution -- and democracy
contained by Counter-Democracy
contained by Deliberative democracy
Main Entry of Democracy -- acculturation of
Main Entry of Democracy -- and China's imprisonment of activists
Main Entry of Democracy -- and deliberation
Main Entry of Democracy -- and Japan (Tokyo 1964 Games)
Main Entry of Democracy -- and

TOPIC NAME


Name
democracy
Scope
Generic
Display
Preferred
Hidden
Normal
Submit
Cancel
Democracy

REVERSED INDEX

Content

and how he demonstrated the value of trust throughout his life. And still another directed students to draw a cartoon illustrating one of the five Community of Caring values. While these activities are not necessarily bad, they highlight the limited nature of human-relations approaches to multicultural education. They are illustrative of the nice ways educators engage diversity and how we typically school youth in niceness. This notion of multicultural education as human relations is a common theme in the literature. Similar to the belief expressed by teachers in the Zion School District, "Advocates of human relations . . . believe the approach needs to be fostered in everyone, and in all schools, to make our democracy work and bring about world peace" (Sleeter and Grant 2003, 81). The goals of human-relations approaches to multicultural education are working toward greater harmony in social relations among all students, encouraging students to learn about cultural differences while respecting others' right to deviate from the norm, creating unity and tolerance among people, and reducing prejudice (see, for example, Gibson 1984; Kincheloe and Steinberg 1997; McLaren 1994; Nieto 2004; Sleeter 1996; Sleeter and Grant 2003). While these goals certainly have some value, a number of critiques have been leveled against human-relations approaches to multicultural education, including that they fail to address the structural nature of inequity, can be assimilationist, implicitly accept the status quo, and are only concerned about diversity when it threatens the perception of harmony and unity (Sleeter and Grant 2003). Thus, while many teachers equate multicultural education with human relations, most scholars in the field argue that a narrow focus on human relations does not go far enough in working for greater educational equity and, ultimately, in challenging whiteness. I would add that a focus on human relations as a manifestation of multicultural education merely reinforces whiteness through its valuing of niceness. Part of being nice means not talking about potentially conflict-laden topics such as discrimination, privilege, and oppression (Boler 2004; Castagno 2008; Howard 1999). In order to "just get along," it

Topics at This Location

- Grant, Carl
- Sleeter, Christine
- democracy
- discrimination
- diversity
- diversity -- discrimination
- dominant
- marginalization
- multicultural(ism)
- oppression
- prejudice
- privilege
- status quo
- systemic

Indexes

Index 1

WHAT'S NEXT?

- Attend Alexandra Provo's and Daniel Lovins' presentations for more information on the project.

- Topic Map applications are generic.

  - Extracting topics can be done from other sources than indexes: databases, spreadsheets, XML, unstructured text, PDF, etc.

- Infoloom has a new version of the Curation Toolkit under development.

CONTACT

- Michel Biezunski
- mb@infoloom.com

- https://www.infoloom.com
- Slides downloadable at https://www.infoloom.com/media/presentations/mb-2018-04-28.pdf