Chris Holstrom. 2019. Moving Towards an Actor-Based Model for Subject Indexing. NASKO, Vol. 7. pp. 120-128.

Chris Holstrom — University of Washington

Moving Towards an Actor-Based Model for Subject Indexing

Abstract

This paper presents a preliminary exploration of an actor-based model for subject indexing, which considers four types of actors: professional indexers, domain experts, casual indexers, and machine algorithms. The paper describes each of the four actors, enumerating differences in approach, training, methodology, priorities, and tools, as well as similarities and historical collaborations between actors. The paper then explores how the actor-based model for subject indexing might serve as a complement to existing models that focus on processes, methods, disciplinary norms, and cultural biases by defining and exploring the following key properties of an actor-based model for subject indexing: 1) actors are the primary drivers of subject indexing work, 2) observing and understanding many types of actors' processes in real-life situations is as valuable as prescribing correct methods for professional subject indexing, and 3) multiple and different types of actors can perform subject analysis work and subject representation work on the same information objects, and these hybrid (multi-actor) approaches to subject indexing are explicitly supported. These key properties suggest that an actor-based model for subject indexing might open new research opportunities and encourage new hybrid and collaborative approaches to knowledge organization.

Introduction

Subject indexing, the process of describing and classifying information objects, comprises two subprocesses: subject analysis and subject representation. Subject analysis (or concept analysis) is the process of determining what an information object is about and what its essential characteristics are. Subject representation maps the output of subject analysis to a knowledge representation or an indexing language. Both parts of subject indexing, subject analysis and subject representation, are always performed by some type of agent or actor. These subject indexing actors all have agency and all act on information objects. Following actor-network theory (Latour, 1996), a subject indexing actor can be a person, object, idea, or process. Also following actor-network theory, these actors can interact with each other directly or through information objects. For example, a machine algorithm might determine what a text is about and then a person might determine how to express the algorithm's determination in an indexing language.

Although actors are integral to subject indexing work, most existing models for subject indexing assume a single type of actor: a professional indexer. These models detail or prescribe processes (ANSI/NISO, 2005), methods (Wilson, 1968), and established disciplinary norms (Cutter, 1904) primarily for professional indexers. This focus on professional indexers reflects the rich history of library science driving the discipline of subject indexing. Adjacent to this rich history is a substantial body of research that describes how cultural bias and other factors affect subject analysis and subject representation work, and how these effects have significant impacts on society (Bowker and Star, 2000). Integral to studies of how culture affects subject indexing is the question of who is doing the work. The actor-based model for subject indexing takes inspiration from these studies of cultural bias in classification and indexing and addresses similar questions using the lens of actors:

- What types of actors perform subject indexing work and what are their defining characteristics?
- How do the approaches, training, methodologies, priorities, and tools of these actors affect their subject indexing work?
- How can our understanding of these actors help us develop new approaches to and a better understanding of subject indexing?

The actor-based model for subject indexing considers these questions by defining four main types of actors: 1) professional indexers, 2) domain experts, 3) casual indexers, and 4) machine algorithms. Like professional indexers, domain experts have a long history of performing subject indexing. More recently, machine algorithms and casual indexers working in folksonomies and other milieus have emerged as subject analysis actors. Focusing on subject indexing actors, especially less explored actors like domain experts, casual indexers, and machines algorithms, present an opportunity to expand our definition and understanding of subject indexing in ways that complement method-, process-, discipline-, or culture-focused models for subject indexing.

Related Work

Some existing research considers the role of actors in subject indexing, often comparing other subject indexing actors to professional indexers. For example, Adler (2009) compares the controlled vocabulary of Library of Congress Subject Headings (LCSH) with user-generated tags in LibraryThing and finds “a disconnect between the language used by people who own these books and the terms authorized by the Library of Congress and assigned by catalogers to describe and organize transgender-themed books.” Kipp (2011) compares how users (casual indexers), authors (domain experts), and professional indexers index journal articles that are available on CiteULike. These actors are shown to use different terminology and orthographic standards and to emphasize or de-emphasize different characteristics, such as geography. Chu and O'Brien (1993) find that novice indexers were able to determine the subject of texts in most scientific fields but were less successful identifying the subject of humanities texts. Hjørland (2002), in developing a domain-analysis approach to information science, explores the general classification knowledge that professional indexers bring to subject indexing and how this knowledge relates to the domain-specific knowledge that domain experts and some professional indexers possess.

Studies also compare how human actors and machine algorithms perform subject indexing tasks or describe how these actors can collaborate. For example, the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015) compares how well humans and machine algorithms detect and classify objects in images and finds that machines have passed humans at identifying objects in images under controlled conditions. The NASA Lexical Dictionary (Silvester et al., 1994) is an early example of machine-aided indexing that uses semantic analysis and a controlled vocabulary to help humans index documents. These machine indexing systems have become more common, powerful, and autonomous as machine learning has progressed rapidly in the past decade. Studies of these systems often implicitly or explicitly compare machine algorithms with professional indexers, based on criteria such as speed and cost of indexing or precision and recall. Golub et al. (2015), noting that "research comparing automatic versus manual indexing is seriously flawed" develop a framework for evaluating machine indexing information retrieval in real-life situations.

These and many other studies of subject indexing and subject indexing actors show that researchers are attuned to the differences between these actors and opportunities to build hybrid and collaborative subject indexing processes, but they do not always explicitly approach their research through the lens of actors or compare types of actors. This paper aims to encourage researchers to examine subject indexing explicitly through the lens of actors and, by extension, encourage new hybrid and collaborative approaches to subject indexing.

Types of Subject Indexing Actors

Four primary types of actor perform subject indexing work: 1) professional indexers, 2) domain experts, 3) casual indexers, and 4) machines. These subject indexing actors all have agency and all act on information objects. While these types of actors are largely discrete, there are two exceptions. First, the same person can act as different actors at different times. For example, a fashion expert might perform subject analysis on the fall collection and then tag travel photographs in a folksonomy. Second, the same person might act as multiple actors at the same time. For example, a biologist might also have training and experience as a professional indexer, or a special collections librarian might have domain expertise and indexing expertise.

Finally, not all actors of a given type are homogeneous. For example, machines can employ different subject analysis algorithms and, as Mai (1999) notes, professional indexers evolve as they gain experience throughout their careers. While heterogeneity within a type of actor suggests that individual actors might be too diverse to characterize as uniform types of actors, we can identify the defining characteristics of each actor type. These defining characteristics help us understand how, for example, domain experts and casual indexers differ and when they might benefit from collaborative and hybrid approaches to subject indexing. The following subsections describe these defining characteristics and compare each of the four subject indexing actors.

Professional Indexers

Professional indexers, who typically have formal training in subject indexing and work in roles such as librarian or taxonomist, are the most studied and most influential type of subject indexing actors. More so than other actors, professional indexers intentionally perform subject analysis for others and across a global scope. Professional indexers do not limit their subject indexing work to their own field as domain experts do or tag documents for personal retrieval as many casual indexers do. Their intent is to describe aboutness and aid retrieval for a broad set of users.

Professional indexers have developed and use standard indexing languages and frameworks, such as Universal Decimal Classification (UDC), LCSH, and Dublin Core Metadata Initiative (DCMI). These sophisticated and generalized systems, some more flexible and some more rigid, typically use controlled vocabularies and enforce some type of taxonomic or ontological structure. Many indexing languages used by professional indexers have detailed classification schemes and indexing and notation rules that are intended for professional indexers with significant training in and experience with subject indexing.

Because of their frequent use of these indexing languages, professional indexers are more likely than domain experts and casual indexers to be influenced by indexing languages when they perform subject analysis. For example, professional indexers might target known or familiar LCSH terms instead of performing an independent subject analysis because they have learned the efficiency of combining these steps. Machine algorithms can be similarly influenced if they are provided a controlled vocabulary while performing textual analysis.

Domain Experts

Domain experts are scholars and practitioners who are extremely knowledgeable within specific fields. Domain experts are particularly capable of analyzing documents within that field and mapping that analysis to indexing languages used in that field. For example, a geneticist is much more capable than others at identifying gene sequences and representing these sequences in the Gene Ontology knowledgebase. Accordingly, we often rely on domain experts to provide definitive answers for the aboutness of complex and specialized documents. Furthermore, professional indexers have historically borrowed from domain experts' scoped indexing languages when building generalized indexing languages and domain-specific and scientific warrant has historically driven scheme change in professional indexing languages.

Unlike professional indexers, who broadly consider the needs of many users, domain experts are more likely to consider the information needs and priorities of other scholars and practitioners in their field. This pragmatic approach undoubtedly helps other domain experts find information, but it might prove less useful for people new to a domain. Similarly, domain experts are somewhat likely to consider indexing languages while performing subject analysis, but less so than professional indexers and most often in cases where indexing languages are established standards in the discipline. These domain specific indexing languages might not interoperate well with global indexing languages and schemes.

Hjørland (2002) argues that professional indexers benefit significantly from also being domain experts, and that this dynamic of professional indexers with deep specialized knowledge is the way forward for the profession in the increasingly large and specialized information environment. This argument is convincing and fits with an actor- based model for subject indexing because the model allows for the same person to function as multiple types of actors. It does, however, complicate the idea that professional indexers work on a global scale while domain experts work on a more local scale. I'd argue that a difference remains, even in Hjørland's framework, in that professional indexers working in a specific domain consider global indexing concerns more readily than do domain experts.

Casual Indexers

Casual indexers typically do not have formal training in subject indexing and, while often avid enthusiasts, typically do not possess the expertise that domain experts possess. Casual indexers are more inclined than other actors to use natural language because they are typically unconcerned with indexing languages or controlled vocabularies. Casual indexers' tendency toward natural language can reflect a community- or user-focused voice in a way that other actors cannot; however, eschewing controlled vocabularies means that casual indexers typically lack consistent terminology when their tags are aggregated.

Casual indexers are often associated with social tagging and folksonomies. In narrow folksonomies (Vander Wal, 2005), a single casual indexer tags each document. In broad folksonomies, many casual indexers tag the same documents, and aboutness is often inferred through community consensus. Casual indexers in broad folksonomies are more likely to use self-specific tags like "todo" or "read_later" that provide minimal subject information (Golder and Huberman, 2006) and have hyper-localized utility that contrasts with the global or domain-specific scope of professional indexers and domain experts, respectively. In both narrow and broad folksonomies, the natural language tags provided by casual indexers are generally not mapped to an indexing language with a controlled vocabulary or semantic relationships. In other words, casual indexers perform subject analysis work but generally do not perform subject representation work.

While casual indexers are often associated with the rise of web folksonomies, casual indexing work has been performed for centuries as categorization or even simply naming or labeling objects. Some examples include untrained volunteers analyzing and categorizing classroom book collections (Holstrom, 2019) and citizen scientists describing instrumentation noise "glitches" (Jackson et al., 2018). Some might argue that subject indexing actors like folk biologists or master gardeners are casual indexers, but their level of expertise, scope of their domain, and size of their intended audience more often aligns these actors with domain experts.

Machine Algorithms

Machine algorithms are unique among subject indexing actors in that they are not humans. Also, while studied extensively in computer science, automatic indexing, machine-aided indexing, machine classification, and related topics, have not been studied as extensively in the field of knowledge organization.

Many knowledge organization scholars view machine algorithms as a tool or extension for other actors to more efficiently implement their subject indexing methods, not as actors in their own right (Foskett, 1996, Svenonious, 2000). This view holds some truth. For example, a machine might simulate the process that a professional indexer uses for subject analysis by looking at the same key parts of documents. Similarly, a machine algorithm might use a decision tree developed by domain experts to analyze documents in that domain.

However, actor-network theory suggests that machine algorithms can be actors, and the actor-based model for subject indexing adopts this approach. Machine algorithms, while influenced by the actors who develop them, have agency of their own and, especially in the case of neural networks, perform subject indexing work differently than humans. Because they function differently, machine algorithms are a particularly interesting type of actor to study and combine with other actors. Machines are the most flexible or amorphous subject indexing actors. Machines algorithms can perform basic categorization or more scientific classification. Machine algorithms can operate on a global, domain-specific, or local scale. Machine algorithms can use controlled or controlled vocabularies. Better understanding machine actors' flexibility and their these differences with human actors represents a significant opportunity to advance knowledge organization research.

A major aspect of this potential lies in the rapid improvement of machine indexing. Artificial intelligence is increasingly effective at analyzing the subjects of texts, photographs, and audio and video documents. As noted previously (Russakovsky et al., 2015), machines have surpassed humans at the subject analysis task of detecting objects in images. Machine algorithms have also evolved from basic reference-counting methods for textual analysis to more sophisticated methods for subject indexing of texts, including semantic analysis, in part because of machine's ability to learn from increasing large training data sets.

Machines use either supervised or unsupervised learning algorithms to perform subject indexing work. Supervised learning algorithms represent a hybrid or collaborative approach to subject indexing because they rely on input and feedback from another actor. Unsupervised learning does not necessarily rely on collaboration with another actor but might rely on a controlled vocabulary supplied by professional indexers or might perform only one of subject analysis or subject representation. Because machines typically perform subject indexing on large sets of data; however, machines often combine subject analysis (or identification) with subject representation (clustering or automatic classification). This unification of subject analysis and subject representation is similar to the approach of many professional indexers and presents opportunities for breaking these steps apart and building hybrid subject indexing methods.

Key Properties of the Actor-Based Model for Subject Indexing

Based on these four actor types, we can begin to see how an actor-based model of subject indexing might differ from and complement existing models, in particular building on culturally attuned models for subject indexing. The actor-based model for subject indexing has the following key properties: 1) actors are the primary drivers of subject indexing work, 2) observing and understanding many types of actors' processes in real-life situations is as valuable as prescribing correct methods for professional subject indexing, and 3) multiple and different types of actors can perform subject analysis work and subject representation work on the same information objects, and these hybrid (multi-actor) approaches to subject indexing are explicitly supported. These key properties suggest that an actor-based model for subject indexing might open new research opportunities and encourage new hybrid and collaborative approaches to knowledge organization.

Actors as Primary Drivers

The fundamental property of the actor-based model for subject indexing is that it considers the diverse approaches and motivations of actors to be the primary drivers and differentiators in subject indexing decisions. This view differs from models that focus on more specific differences in subject indexing, such as knowledge representations (i.e. ontology versus thesaurus), indexing languages (DDC versus LCSH), approaches (enumerative versus synthetic), or techniques (purposive versus appeal to unity).

There is a rich literature about how bias affects subject analysis and representation, and how these biases can shape societies (Bowker and Star, 2000). Like cultural biases, the inherent biases in approach and motivation of different actors can significantly affect subject indexing decisions. For example, professional indexers follow prescribed processes and have a broad set of users in mind while casual indexers most often have themselves in mind. These differences can significantly affect subject indexing and represent a rich opportunity for comparative studies that may produce findings similar in scope to those found by researchers studying subject indexing through the lens of cultural and institutional bias. For example, if we recognize and embrace the differences between actors and their methodologies and tools, we might find that machines are better or worse at identifying emergent topics based on their approaches to literary warrant. Or we might find that domain experts develop knowledge representation formalisms that could be applied generally to global indexes. Or we might find that casual indexers using community authored knowledge organization systems can establish a channel for minority voices in a way that professional indexing systems do not support (Holstrom, 2018).

Observation in Real-Life Scenarios

An actor-based model focuses on understanding and describing how different actors approach subject indexing in real-life situations. This approach contrasts with much traditional research on professional indexers in that it does not attempt to arrive at "correct" approaches or methodologies for subject indexing. Instead, all four subject indexing actors are equally privileged and present equal opportunity for observing novel and useful subject indexing work. This model, then, emphasizes description, not prescription, and does not aim to guide subject indexing actors to the one true way to perform subject indexing work.

Because the actor-based model for subject indexing considers four different types of actors and aims to observe instead of judge whether these actors' approaches are right or wrong, research using this model might identify specific behaviors or processes that these actors exhibit. These behaviors and processes, particularly those observed in less studied actors, can contribute to subject indexing as a whole, much in the way that observing social tagging behaviors helps us identify emergent vocabulary or observing machine learning clustering helps us understand new relationships between subjects or observing domain experts' indexing choices have informed global indexing languages. More opportunities to borrow ideas and practices from other actors are likely to present themselves if we simply observe—and many of those opportunities for borrowing might arise from observing what professional indexers actually do, not what is prescribed.

Hybrid and Collaborative Subject Indexing

An actor-based model for subject indexing, inspired by Langridge (1989), advocates that subject analysis be separate from subject representation and explicitly supports hybrid approaches to subject indexing, for example casual indexers doing subject analysis and machines aligning that subject analysis with an indexing language. Keeping these steps separate is particularly well-suited to an actor-based model because of the opportunity to combine the strengths of one actor in subject analysis with the strengths of a different actor in representation. The many possible multi-actor combinations suggest a particularly rich set of research opportunities. For example, one might study how different actors could use folksonomy data to build structured indexing languages.

An actor-based model also expressly supports collaborative (multi-actor) approaches to subject indexing. For example, domain experts and professional indexers might work together on subject representation or casual indexers might use machine-suggested terms while performing subject analysis. By explicitly defining and better understanding actor types, we can better understand and more intentionally encourage collaboration between actors. We might also better understand “double actors” like special collection librarians, where a single person functions as a professional indexer and a domain expert, by isolating when this person uses each actor type's processes or methodologies. Finally, we might better understand the relationships between human actors and machine actors, their comparative strengths for specific tasks, and when machine actors should adopt methods and processes from specific human actors. If so, we can design more effective machine-aided indexing processes that consider subject indexing best practices and the rich research history in knowledge organization.

Conclusion

This paper proposed and explored an actor-based model for subject indexing intended to complement existing process-, method-, discipline-, and culture-focused models. Studying subject indexing through the lens of actors—professional indexers, domain experts, casual indexers, and machines—might open new research opportunities and encourage new hybrid and collaborative approaches to knowledge organization. These opportunities and hybrid approaches are particularly unexplored for machine algorithms, which represent the largest opportunity for knowledge organization research to use an actor-based model to develop a richer understanding of and new best practices in subject indexing.

References

Adler, M. 2009. Transcending library catalogs: A comparative study of controlled terms in Library of Congress Subject Headings and user-generated tags in LibraryThing for transgender books. *Journal of Web Librarianship* 3(4): 309-331.

ANSI/NISO. 2005. Guidelines for the construction, format, and management of monolingual controlled vocabularies. Bethesda, MD: National Information Standards Organization.

Bowker, G.C., and S.L. Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

Chu, C.M., and A. O'Brien. 1993. Subject analysis: the critical first stage in indexing. *Journal of Information Science* 19(6): 439-454.

Cutter, C.A. 1904. *Rules for a dictionary catalog*. Washington, DC: Government Printing Office.

Foskett, A.C. 1996. *The Subject Approach to Information*, fifth edition. London: Library Association Publishing.

Golder, S.A., and B.A. Huberman. 2006. Usage patterns of collaborative tagging systems. *Journal of information science* 32(2): 198-208.

Golub, K., D. Soergel, G. Buchanan, D. Tudhope, M. Lykke, and D. Hiom. 2016. A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology* 67(1): 3-16.

Hjørland, B. 2002. Domain analysis in information science: eleven approaches– traditional as well as innovative. *Journal of Documentation* 58(4): 422-462.

Holstrom, Chris. 2019. 'Is This a Chapter Book?': Parent-Involved Categorization in a Kindergarten Classroom Book Collection. *Cataloging & Classification Quarterly* 57(2): 1-14.

Holstrom, Chris. 2018. Local Authorial Voice and Global Authorial Voice in Community-Authored Knowledge Organization Systems. *SIG/CR Workshop, ASIS&T 2018*. Vancouver, B.C.

Jackson, C., K. Crowston, C. Østerlund, and M. Harandi. (2018). Folksonomies to support coordination and coordination of folksonomies. *Computer Supported Cooperative Work (CSCW)* 27(3-6): 647-678.

Kipp, M. E. 2011. "User, author and professional indexing in context: An exploration of tagging practices on CiteULike/Le contexte de l'indexation des usagers, des créateurs et des professionnels: Une exploration des pratiques d'étiquetage social sur CiteULike." Canadian Journal of Information and Library Science, 35(1): 17-48.

Langridge, D.W. 1989. *Subject analysis: principles and procedures*. London; New York: Bowker-Saur.

Latour, B. 1996. On actor-network theory: A few clarifications. *Soziale welt*, 369-381.

Mai, J. 1999. Deconstructing the indexing process. In *Advances in Librarianship* (pp. 269-298). Emerald Group Publishing Limited.

Russakovsky, O., H.S. Jia Deng, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3): 211-252.

Svenonious, E. 2000. *The Intellectual Foundation of Information Organization*. MIT Press.

Silvester, J.P., M.T. Genuardi, and P.H. Klingbiel. 1994. Machine-aided indexing at NASA. *Information Processing & Management* 30(5): 631-645.

Vander Wal, T. 2005. Explaining and showing broad and narrow folksonomies. http://www.vanderwal.net/random/entrysel.php?blog=1635

Wilson, P. and D.W. Koepp. 1968. *Two kinds of power: An essay on bibliographical control*. Univ of California Press.