ToDI: A Taxonomy of Derived Indices

Maria Joseph Israel(✉), Navid Shaghaghi, and Ahmed Amer

Ethical, Pragmatic, and Intelligent Computing (EPIC) Laboratory,
Santa Clara University, Santa Clara, CA, USA
{misrael,nshaghaghi,aamer}@scu.edu

Abstract. Advancements in digital technology have eased the process of gathering, generating, and altering digital data at large scale. The sheer scale of the data necessitates the development and use of smaller secondary data structured as 'indices,' which are typically used to locate desired subsets of the original data, thereby speeding up data referencing and retrieval operations. Many variants of such indices exist in today's database systems, and the subject of their design is well investigated by computer scientists. However, indices are examples of data derived from existing data; and the implications of such derived indices, as well as indices derived from other indices, pose problems that require careful ethical analysis. But before being able to thoroughly discuss the full nature of such problems, let alone analyze their ethical implications, an appropriate and complete vocabulary in the form of a robust taxonomy for defining and describing the myriad variations of derived indices and their nuances is needed. This paper therefore introduces a novel taxonomy of derived indices that can be used to identify, characterise, and differentiate derived indices.

Keywords: Data indices · Index derivation · Metadata hierarchy · Referential data · Taxonomy

1 Motivation

Advancements in digital technology have eased the process of gathering, generating, and altering digital data at large scales. As a result, publishing, intellectual property, or moderation of data raises questions of proper attribution, ownership, and fair use. These questions are further complicated when dealing with data that refers to other data, *i.e.*, referential data, metadata, indices, *etc.* and it would be impossible to speak rigorously and meaningfully about solutions to problems in this space (which will only increase as the volume of digital corpora increases) without a clear taxonomy of the different dimensions of metadata and its derivation. To address this issue, we have developed a taxonomy to assess the nature and shape of all the different forms and collections that such referential data can take. At the time of this writing, to the best of our knowledge, the Taxonomy of Derived Indices (ToDI) presented in this work is the first of

N. Shaghaghi et al. (Eds.): INTETAIN 2020, LNICST 377, pp. 43–61, 2021.
https://doi.org/10.1007/978-3-030-76426-5_4

its kind to provide a clear and rigorous vocabulary for describing instances of derived or referential data in the form of indices.

The focus of this paper is not so much on the specific implementations of different index types, but rather on the basic variations in the relationship between indices (and similar metadata) and the data from which it is derived. We make some basic assumptions regarding data and metadata, but attempt to keep our definitions as broad and inclusive as possible.

The key question is whether a datum is solely a primary representation of data, or whether it was a result of derivation from existing indices. This brings about complex questions regarding originality, authorship, ownership/intellectual property, attribution, responsibility, and privacy. We cannot be confident in rigorous analyses of such questions, let alone ethical evaluation, if we cannot describe clearly the relationship of data to its derivatives. The presented taxonomy here is an attempt at laying the foundation for others to achieve all of that.

The rest of this paper is structured as follows: Sect. 2 introduces basic terminology such as data, metadata, and index, as way of providing background for discussing other key terms of the taxonomy and Sect. 3 details related work. Section 4 describes the structure of the proposed taxonomy and key descriptors and Sect. 5 explores the future work on ethical frameworks which this taxonomy contributes to. And lastly, Sect. 6 provides some closing remarks.

2 Background

This section provides a brief overview of some basic but important terms and terminologies such as data, metadata, and index, in order to set the background for describing a taxonomy as these terms are implicitly referred to in describing the taxonomy.

2.1 Data

We live in an era of big-data with unfathomable amounts of data everywhere. Data is often associated with electronic data produced by and stored on the informational technology innovations of the twentieth century, though the use of the term "data" is much older. Datum, a singular element of data, can be understood as a basic unit of information. When it is aggregated it forms the bedrock of knowledge. Data differs from information and knowledge, in the sense that it is discrete and descriptive, and is dependent on interpretation for it to hold meaning (and thus become information). To this, Gitelman prepends other characteristics such as abstractness, aggregativeness, and graphicalness [8] when he discusses the origins and economy of modern data and offers the following rationale: Data is abstract, because it requires material expression. Data is aggregative because it can pile up into larger datasets or be reduced into smaller datasets. Data is graphical because it can be visualized by diagrams and graphs as part of an explanation of one's argument as well as to inform us of what we already know or ought to know, and drive us either to action or inaction depending on circumstances (such as data-driven government policies, financial investments, scientific experiments, medical and health practices, ecological claims, personal decisions, *etc.*). Data is therefore a symbolic, aggregatable, abstraction of the concepts that we interpret it to represent.

While the underlying idea of data remains the same, the relationship between data, data creation, and data consumption has radically changed. This is a result of our increased ability to generate and process ever greater quantities of data. That is to say, we have moved from passive data consumers to active data creators, as today's information technology makes it easy and relatively cheap or even free to create, manipulate, accumulate, store, and transmit data at high speed and large quantities. This characteristic nature of data has also reversed the role of data from being subject to us, waiting upon our interpretation of small amounts of data, to we being subject to the data, given that every click and every move has the potential to count for something for someone somewhere, meaning that every individual can produce such vast quantities of data, that no individual can reasonably consume or process it all. For example, the Google Search Engine collects information on user's interests and behaviors [19], Amazon's Alexa gathers user's personal information and interactions [15], Facebook tracks user's social friends and their activities [4], just to name a few. There is a radical shift in contemporary conception of data and its economy on everyday life activities and interactions. This raises complex questions of naming such data for what they are, let alone ethical questions raised by the use and abuse of such data. Descriptive language for all this data is complicated based on activities that alter data creation, collection, and its subsequent custodianship or ownership.

2.2 Metadata

Closely related to data is the concept of metadata that is gaining currency in the field of big-data and online social media. Metadata, generally defined as "data about data" [11], describes additional information about data and its movement and modification among individuals, organizations, disciplines, and machines. Paradigmatic metadata use cases include library catalogs, table of contents, and inventories. With the advent of information technologies, metadata is gaining popularity and is widely used both by information professionals in cataloging, classifying, and indexing data, and by information consumers in social networks' tag clouds, folksonomies, photo captions, and social bookmarks. Therefore, metadata can be understood as the sum total of value-added information/description of any information object's content, context, and structure at any level (*i.e.*, individual, aggregate, or system) [7]. Often metadata elements are structured as explicit information, compiled in standardized categories, and tightly controlled vocabularies in order to establish common ground for researchers and users [13] and to enable authoritative, inter-operable, scalable, achievable, preservable record-keeping systems [7]. For example, archival and museum metadata contain information such as indices, abstracts, and bibliographic records adhering to data content standards like Anglo-American Cataloging Rules (AACR), Machine-Readable Cataloging (MARC), and Library of Congress Subject Heading (LCSH). Scientific metadata datasets specific to Earth System Science comply with, among many others, the Open Archival Information System (OASIS) Reference Model and the Network Common Data Form (NetCDF). With the advent of RDF and semantic technologies, rich metadata are not only manually generated/curated, but automated through metadata mining, metadata harvesting, and web crawling.

However, the origin of metadata goes back to when humans began using language to communicate their feelings, thoughts, and plans. One of the earliest recorded examples of metadata is the use of meta tags, at the great library of Alexandria in 280 C.E., attached to the end of each scroll containing information like title, subject, and author, assisting library users to search through the scrolls without having to unroll all the scrolls and to return a scroll to its right position after usage [5]. This practice evolved into standard library catalogs in the eighteenth and nineteenth centuries and evolved into online catalogs with the emergence of information technology in the late twentieth century [5]. The modern metadata requirement was introduced via meta language by MIT's Stuart McIntosh and David Griffel [9] and metadata standards by the International Press Telecommunications Council [20]. In the early twenty-first century, the National Information Standards Organization introduced metadata for libraries [17] and Adobe's Extensible Metadata Platform (XMP) made use of XML and RDF for metadata representation [2]. Eventually metadata became a pivotal tool for marketing by companies (like Google and Amazon) and leveraged machine learning techniques to automate metadata mining and harvesting [18]. The usage of such vast stores of metadata has thus raised countless privacy and intellectual property concerns to levels that have not yet been fully regulated nor even understood. One attempt to counter the misuse of metadata collection and its management, for instance, the European Union introduced the General Data Protection Regulation (GDPR) compliance law [21].

2.3 Index

Another term that needs to be defined is index which is now generally understood as a mechanism to optimise access to certain data records within a file. The modern use of index traces back to the mid 18th century [22]. In today's digital database context, a database index is a structure similar to a book's table of content utilized for quick data retrieval operations achieved by minimizing table traversal and maximizing performance [1]. In other words, an index is associated with a table and used to efficiently locate data without having to investigate every row in a database table. An index is a routine way of maximizing performance of the databases, at the cost of extra replicas of data.

An index file consists of index entries of the form search key value and pointers to blocks of data in a data file. There are two types of indices: ordered indexes and hash indexes. In the former, search keys are stored in a sorted order and in the latter, the search keys are distributed uniformly across buckets/slots using a hash function. The hash function $f(k, n)$ denotes the key k and the number of buckets n, and maps the key k to the corresponding bucket of the hash index. There are many more types of indices such as primary, secondary, clustering, dynamic, B-Tree, and bitmap indices available to achieve various performance and storage requirements. However, ToDI is not dependent in the 'form' of the index as listed above, but rather in the 'nature' of the index. Section 4 explicates the assumptions, structure, and descriptors of the new proposed taxonomy.

3 Related Work

Given the new ground explored in this work, it is insufficient to focus on a single area of related works; thus a discussion of relevant historical usage and contexts is included.

Before the modern adoption of the term index, tracing back to its classical usage in the mid 18th century, the term index had the same indication or meaning as words like: Table, Register, Calendar, Summary, and Syllabus, and conveyed the meaning of a discoverer, discloser, informer, catalogue or list, inscription, title of a book, and the fore or index-finger [22]. Index was meant to be an indicator, pointing out the position of the desired information. As observed by Wheatley [22], some early usages of (English) indices as tables of references arranged in alphabetical order placed either at the beginning or end of a book, include: Biblical concordances (*e.g.*, an earliest one dedicated to Edward VI by John Marbec in 1550); indices to publications of societies; indices of atlases; indices to statutes; indices of parliamentary papers; indices to the journals of the houses of lords and commons; and indices of periodicals. In other words, an index of the historical uses and forms of indices would not be short.

In more modern usage, within the domain of digital databases and storage systems, an index helps retrieve data from a database system quickly by minimizing data structure traversal and physical access demands, thereby maximizing performance [1]. As a result, most technical researches on indices have looked at their use for performance improvement in various contexts. Several studies have addressed factors concerning the speed and storage efficiency of various index structures like T-Trees, B-Trees, and *etc.*, for main memory databases [3,14] as well as disk-based database systems [10,12,16]. There are also works that focused on the sequential/single dimensional and multidimensional features of index structures in relational database systems [6]. Some examples of the former include: dense index, sparse index, multilevel index, secondary index, B-Tree index, and hash tables; and examples of the latter include hash-like structure based index, tree-like structure based index (k-d tree, Quad tree, R-tree, *etc.*), and bitmap indices [6]. These classifications of indices deal with data storage and retrieval performances of mostly relational database systems. Unlike such a technical focus on index form and performance, the proposed taxonomy presented in this paper is technology agnostic, and is therefore not restricted to any specific realization or application. It also deals particularly with the nature of derived indices, as distinct from base indices.

4 Taxonomy: Assumptions, Structure, and Derivatives of Referential Data

For the design and development of ToDI, the life cycle of data is assumed to include creation/capture, modification, transmission, storage, update, access, archive, restore, and delete. The focus of this research paper is on the modification and access of data by a user or a software agent that operates on it. Naturally, any modification and transformation of original data brings forth another set of data, a derived data. Therefore, derived data is defined as data that has come about either partially or fully from other data sources. Though a data source can be literally anything, for the purpose of this paper, the (main) source of referential data is the actual data without which an index or other related metadata could not exist. It is therefore the primary corpus of digital data without which any derived metadata is not defined, as it would have no context. Depending on the instance of derived data the source may vary. For example, when the first instance of any index or metadata is further abstracted, then the source of the newly derived data is the immediate metadata, not the original data upon which it was defined. Similarly, referential data is defined as data that performs the functions of pointers, indicating where or how to reference other data.

This is different from the main content of a data repository. For example, chapters in a book can be considered as the main content of the book, whereas an index at the end of the book, by its very nature, does not form the primary data of the book, but rather helps reference content/concepts in the main body of the book. In the database domain, a data repository is the actual data, but indices are typically pointers to the main data, structured in some useful manner. Similar to a book's table of contents, some indices can be a part of the book itself, while others, like a bibliographical index, could typically be seen as a separate entity. Either way, there is a logical distinction between the base data, and the data that is derived from it (*e.g.*, a table of contents or bibliographic index), regardless of whether that derived data is typically found along, or apart from, the data from which it was derived.

4.1 Taxonomy Assumptions

The following assumptions determine the scope of the taxonomy:

A1. The types of data that are considered in this paper are indices, also known as referential data, metadata, or pointers/derivatives of original data.

A2. The type of classified index in this taxonomy refers to the nature of the derived indices depicting implicit relationships between original data and derived data, and not based on any data storage and retrieval performance factors.

A3. The proposed taxonomy is generic and tech-agnostic in nature, meaning that it is not restricted to any specific type of data or database architecture. In other words, it is applicable to all heterogeneous data types and database systems.

4.2 Taxonomy of Derived Indices (ToDI)

The different types of indices are categorized based on the nature of their derivation. Figure 1 illustrates the taxonomy of derived indices graphically and Table 1 summarises key characteristics of each of the derived indices. This section describes the taxonomy and presents a case for each type of the derived indices by discussing the logic behind the structure of the taxonomy. Each node of the taxonomy is illustrated with examples as appropriate.

The taxonomy is based on the premise that indices ares a kind of metadata and metadata is itself data. Moreover, like raw data, metadata can be generated directly from data or from modifying, combining, or altering existing metadata. The taxonomy follows a successive hierarchical refinement approach in simplifying the classification of child nodes in the taxonomy. Different taxonomies can be created depending on the different premises. However, the presented taxonomy is the first of its kind and even though it may require further refinement by including finer types to make it more comprehensive, this does not diminish its usefulness and significance.

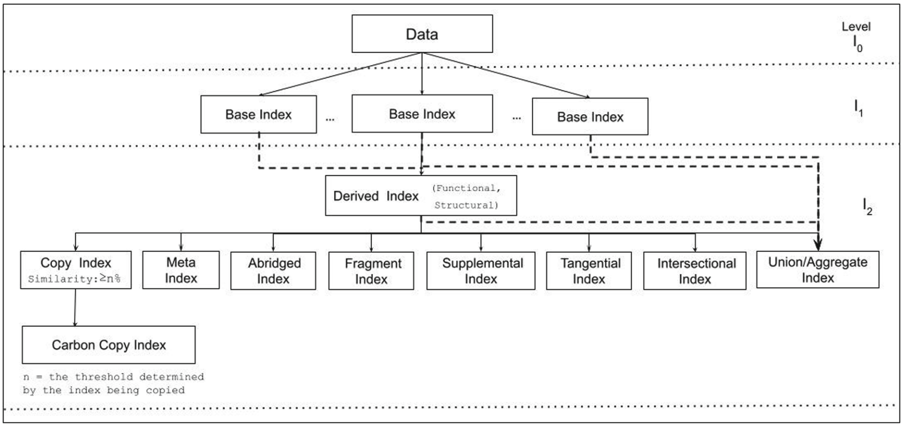

Fig. 1. A Taxonomy of Derived Indices (ToDI)

The root of the taxonomy is a data node which represents a collection of raw data. An example of a dataset which will be used throughout this section is the information housed within a book. A book is usually defined as a set of sheets or pages comprising of sentences containing fictional or nonfictional information that is organized in sections, chapters, and paragraphs.

Table 1. Derived indices: types, definitions, and examples

Index type	Description	Example reference
Base Index	An immediate index of a data set	BI
Derived Index	An index that is derived from an existing index	DI
Functional Derivation	Refers to the functional aspect of the index	FD
Structural Derivation	Refers to the structural aspect of the index	SD
Copy Index	An index that duplicates a portion of an existing index (beyond a threshold)	DI0
Carbon Copy Index	An index that duplicates a hundred percent of an existing index	DI0
Meta Index	An index of an index	DI1
Abridged Index	A briefer version of an existing index	DI2
Fragment Index	An incomplete portion of an index	DI3
Supplemental Index	An index which supplements an existing index	DI4
Tangential Index	An index which indexes related data to data already indexed in an existing index	DI5
Inter-sectional Index	An index which combines parts of existing indices into a single index	DI6
Union/Aggregate Index	An index which combines 2 or more indices in their entirety into a single index	DI7

Base Index (BI). A base index is an immediate index of a data set that is derived directly from the data it indexes without using any other metadata or indices about the data it indexes. All the indices described in the background and related works sections above are examples of base indices. An example base index for a book could be the table of contents included in the book, which is used to index the location where each chapter or section of the presented data in the book can be found.

Derived Index (DI). Derived indices are the result of indexing an existing index or the duplicating, summarizing, supplementing, and/or combining of several indices (be they base or derived) to form a new index of a data set. Derived indices may be the result of structural or functional derivation, or a combination of the two.

Functional Derivation (FD) of an index refers to the derivation of the functionality and purpose of the index that is being derived from. The result of a functional derivation of an index is a derived index that in some way indexes or duplicates, summarizes, supplements, and/or combines existing indices' functionality. Meaning the derived index copies, limits, or enhances the indexing capability of an existing index's referential purpose and functionality regardless of whether it is structured similar to or different from the index or indices it is derived from.

Structural Derivation (SD) on the other hand, refers to indexing or duplicating, summarizing, supplementing, or combining an existing index in precisely the same way that the existing index indexes the data. Meaning the derived index in some way mimics the look and feel (a.k.a. structure) of the index or indices it is derived from. This type of derivation has the potential to violate the copyright and patent protections of the original index if any exist. But this discussion is beyond the scope of this paper and will be explored in a future paper by the authors.

ToDI distinguishes a set of eight basic types of derived indices ($DI0..DI7$): Copy & Carbon Copy, meta, abridged, supplemental, tangential, intersectional, and union/aggregate indices. Each of which is explored below:

DI0. Copy & Carbon Copy Indices. A copy index, is a derived index that is created through any process that duplicates (copies) some or all of an existing index's functionality, structure, or both; but which need not result in a perfect structural and/or functional duplicate of the original index. The amount of the original index which needs to be duplicated in the derived index before that derived index is considered to be a copy index is dependent on the context of the original data and differs from case to case. A copy index may vary from a carbon copy index to a certain degree of cloning of an index which could be distinguished by a degree of (intentional) similarity of the index. In other words, the degree of similarity of a copied index with the original index is a range between some lower bound and 100%. The lower bound is dependent on the data and the purpose of the index being created and thus varies from index to index. A 100% copy is a carbon copy index, similar to the olden days dittoed blue carbon copy paper. That is to say, since the degree of similarity of a copied index to the original index is defined within a range, when the degree of similarity is exactly at 100% of its function and/or structure, then the copy is a carbon copy index. For this reason, a carbon copy index can also be called a clone index.

A copy (and as an extension a carbon copy or clone) index can be a functional copy and/or structural copy, based on the properties of the derived index from which a copy is made. A functional copy of an index functions similarly to the original index, but its design and specific purpose may be different. It may differ in structure, but aid the same functions as the original index. Alternatively, it could be a duplicate of a partial or complete structure, making it a structural copy, that in turn need not be used for the same functions.

An example of a carbon copy or clone index would be a table of contents copied from a book. Since a publisher may have produced a table of contents included with the text, it is possible for another party (such as an online bookseller) to simply copy all of the chapter title and page number information from that table of contents, hence producing a structural carbon copy of that index, or replicating the same work to produce an informationally equivalent but structurally different table of contents, thus producing a functional carbon copy of that index. The former is a base index created by a third party (publisher) from the original data (from the author), while the latter is a derived index.

A more complex example of a structural copy index could be using a traditional book index to produce a table of keyword page occurrence counts. The data from a traditional word index, listing the pages on which the word occurs, would be copied and used to serve a different function, which in this case could be to offer a count of the number of pages on which the word occurs. This is an example of a different functional use of structurally identical data. On the other hand, an example of a functional copy index might be a data structure listing the locations of words in a book organized in ascending order of occurrence (instead of alphabetically), or arranged in a sequence dependent on a numerical hash of the word letters thereby creating a hash table of the words, as opposed to the more traditional alphabetical listing that would be found in a book. Such an index would be derived from the existing index and used to serve the identical function as the existing index but with a fundamentally different structure.

As this example also demonstrates, it is possible for a derived index (such as a carbon copy index) to be indistinguishable structurally and/or functionally from the base index it is derived from. Therefore, it is not the content or functionality of the index itself but rather the manner of its creation that defines its characteristics. It is possible for two people to independently come up with structurally similar (relatively unlikely) or functionally similar (more likely) copies of a derived index. What then makes such indices carbon copy indices is the manner in which these indices are created. If they are created independently then they are both base indices of the same data but if one is created off of (*i.e.*, copied from) another then the one that was produced via copying is a derived copy index of the other index.

DI1. Meta Index. A meta index is an index of an index which in turn refers to the data being indexed. An index of indices occurs in many contexts, and is not directly based on the original data. Sticking with the book analogy, a bibliography of books and book chapters in a particular subject area is effectively a meta index of indices. But such an index, built upon data that is a set of existing indices, need not simply be a higher-level implementation of the same functionality as its underlying indices. One usage domain for such an index, that is not concerned with locating individual indices but rather quite the opposite, it's obfuscation, is the removal of "personally identifiable data elements" in an index of a users' data records. The use of a meta-index as an additional layer of indirection, doubly distances the user from the original data records, and thereby is a means of obfuscation rather than location of data. This is an implicit goal in building general purpose recommendation systems based on individual tracking of user behavior. The result is a broader indexing of indices, intended to provide a wider view of the original data, while simultaneously obfuscating individual behaviors (indices) that contributed to that broader view.

DI2. Abridged Index. An abridged index is an index that focuses on certain select aspects/sections of an existing index. An example of an abridged index is seen in books which poses two indices: one that only lists the chapters and another which lists the chapters along with every section and subsection of each chapter. If the more detailed index is for example assumed to be the base index, then the chapters-only index is a shortening (abridgement) of that index.

DI3. Fragment Index. A fragment index is an incomplete portion of an index, which differs from an abridged index in that it was not necessarily constructed deliberately as an abridged index, and thus is inconsistent in its missing components. An index is therefore fragmentary if it is created through an inconsistent or unplanned process of abridgement. An example of such an index is a table of contents of a book which is missing a page. This table of contents thus might include an incomplete selection of subsections for one chapter, while being complete for all the preceding and following chapters. Any salvaged part of an index of an ancient book is thus also an example of a fragmented index.

DI4. Supplemental Index. A supplemental index is an index that supplements an existing index. In other words, it is a value-added index as it adds additional information to the existing index. In terms of our taxonomy, this would differ from the other derivations in that it is an index that merges additional data that is not inferrable from the original data or the index from which it was derived. An example of a supplemental index in a book is seen in various paper copy textbooks which include several additional (not included on paper) chapters on either an included digital media such as a disk or the textbook's/author's/publisher's website. The added chapters are accompanied with new index entries which supplement the existing index.

DI5. Tangential Index. A tangential index is an index which indexes related data to data already indexed in an existing index. Unlike the Supplemental index, this type of index merely links data that already exists in the original data corpus with existing index data. Since it does not add data from an external source, it is therefore tangential, touching on more of the existing data, but not supplementing it with additional information that could not be derived solely, albeit with extra work, from the existing index. For example, a book word index, listing the pages upon which a word occurs, can be enhanced with a tangential index that adds the line number or numbers in that page wherein the word occurs.

DI6. Intersectional Index. An intersectional index is an index that combines certain parts of other indices into a single index. For example, a table of contents in a book that is an editorial work which brings together parts of several books and lists several chapters, sections, and subsections of each in one uniform index, is an intersectional index.

DI7. Aggregate/Union Index. An aggregate/union index is one that combines two or more existing indices in their entirety into a single index. Any book series which after completion is then republished in a single volume edition may contain a master index which brings together the indices of each of the books together in a single index. That union of indices is an aggregate or union index and not an intersectional index because it includes all of the individual book indices in their entirety.

The taxonomy described is not limited to books and applies equally well to all indices in their varied forms and incarnations. Furthermore, these categories are not mutually exclusive as combinations are easily possible. For example, a separate tangential page occurrence index for all words that are included in each novel written by the Brontë sisters (Anne, Emily and Charlotte) could be created. Then these indices could be aggregated into a master union index which is used as a basis for a supplemental derived index that lists synonyms and antonyms for each word in the novels to produce a thesaurus specific to this body of work. In other words, this would produce a literal Brontë-saurus[1] which is a supplemental aggregate index of tangential indices to each of the table of contents (base indices) in Brontë sisters' novels.

4.3 Taxonomy's Structure and Hierarchy

ToDI deals with referential (anything to which we can refer) data, which by its nature is data that refers to anything, but typically that to which it refers is itself data. This means that referential data is relative to some starting point, that may well be arbitrary. An example could be a single digital data word (or a single binary datum, *i.e.*, a bit) which represents the smallest, simplest, unit of data to which a reference can be made. If debating the nature of scholarly publication indices, it would be reasonable to call a single published paper the atomic datum in such a context. If, however, we are talking about indices built upon a textual database like a book, then the smallest item to which one can refer can be an individual character or symbol within the book. Where one chooses to start, answers the following question of what the lowest level of granularity is within the system, *i.e.*, level 0 or the source data itself, as opposed to any reference to it. If one were to build a higher-level index, not of papers, but of existing scholarly indices, and cared not for any individual paper, then level

[1] We use this particular example to point out that such an index can have surprisingly comical uses. This example, and terrible pun, was inspired by the works of British writer and comedian John Finnemore.

zero would be a single index, with its data items being the entries of the index (which may seem clearly referential data, but would not be considered level 0 if one does not ever actually refer to any individual papers within the system). A table of contents of a conference proceedings is a simple example of a level 1 index, which in this case does not make use of any of the referential data that could be decoded from the papers (*i.e.*, the bibliographies and citations).

In the proposed taxonomy, there are many levels of indices based on the nature of the derived index/indices. To determine the order of these different types of indices, the formula: $I_{(i+1)}$ can be used. Here, I indicates the index and $i+1$ indicates the current order of the index which is the result of incrementing the order of I's parent index (i) by 1 to arrive at the order of the index. This simply means that the raw data is I_0, the first (base) index of the raw data is denoted by $I_{0+1} = I_1$ to indicate that it is of the first order level, and a derived index from it is denoted by $I_{i+1} = I_2$ to indicate that it is of the second order level, and so on ad infinitum. That is, any index that is derived from index order i, is a derived index of the order $i+1$. The different types of indices and their relative order is depicted in Fig. 1 and enumerated in Table 1.

5 Future Work

Given the advancement in digital technology, it is not only easy to generate and alter data at large scale, it is also feasible to add, modify, and delete underlying metadata that point to the location of original data/information. Such metadata can be of index types which could be further abstracted to distance/alter from the original data, or otherwise transcribed, translated, or modified to varying degrees from the original data and the original index. Descriptive language for all this data, and the nature of their potential links and associations, is complicated based on activities that automated data creation, collection, and alteration, in such a manner as to bring into question its rightful custodianship, ownership, or even authorship. This raises ethical questions concerning the use and abuse of such data, especially issues of proper attribution, ownership, and fair use of such digital data and derivatives thereof. While there are efforts to address the ethical and legal dimensions of these issues, a consensus regarding their desired properties is yet to emerge. In this context, this new taxonomy can be used to guide, compare, or differentiate, the different ethical and legal frameworks that have been, and may yet be, proposed. This could be very beneficial for the clear communication of arguments and proposals regarding ownership and attribution of metadata for instance.

The ethical question "Is data about you yours, or should it be?" is an important one for discussing the nature of ownership of data in general, but any such discussion is incomplete if it does not address derivative and referential data as well. For example, how would we describe the data in the following scenarios: Is data about person X, when X's personally identifiable information is included within the metadata, considered the data of person X? What if the personally identifiable data was not included in the corpus, but could be inferred by processing derivative metadata? Would the anonymized data be of person X with or without the metadata? Is the metadata considered to be about person X if it alone would not identify the person? To address these questions, and to coherently discuss questions of identification, in addition to attribution, custodianship, and ownership, the vocabulary provided by the proposed taxonomy of derived indices would become very useful. Further exploration of this space, including comparative analysis of case studies of intellectual property arguments involving metadata, and privacy arguments surrounding the use of metadata, is warranted. Such explorations would be direct applications of the presented taxonomy, and would be expected to demonstrate its usefulness for providing a coherent common language when describing data that described data.

6 Concluding Remarks

This paper delineates a novel taxonomy of derived indices and explains its potential usefulness in exploring ethical questions surrounding metadata. The motivation for the research is the confluence of big data, specifically the increasing ability to manipulate and manage ever-larger datasets, which in turns aids the increasingly easy modification, abstraction, and duplication or recreation of metadata. Such data, whether ubiquitously captured from users' interactions on social media, or mined from ever-growing logs of transactions and activities, is increasingly vast, but so is our ability to generate more useful representations, summaries, and references to the data. So far the focus of discussion has naturally been concentrated on the original preserved data, and less on the derived metadata (*i.e.*, indices). The presented taxonomy specifically focuses on this easily overlooked form of data, which with ever-larger datasets becomes increasingly valuable, but which also is – by its very nature – a derivation with varying degrees, of the original data.

The objective was therefore to present a taxonomy of derived indices that provides a basis for systematically understanding the complexity of different forms of referential metadata, and thereby introducing a useful vocabulary to discuss them. And the proposed taxonomy is an initial offering which may well require further iterations of refinement and development to ensure that it is comprehensive and complete.

Acknowledgement. Many thanks are due to the departments of Mathematics and Computer Science (MCS) and Computer Science and Engineering (CSEN) for their continued support of the project. And to reviewers of ToDI's drafts for their helpful comments without which this work could not have been improved.

A Appendix

A.1 Taxonomy and Associated Descriptors

Term	Description
Term Name	Base Index
Label	Base Index
Definition	A base index is an immediate index of a data set
Comment	A base index is different from other derived indices in that it is base/source of all other derived index (indices). All the indices described in the related work section of this paper are examples of base indices. An example of a base index is for a book could be the table of contents, which is used to index the location where each chapter or section of the presented data in the book can be found
Type of Term	Base index

Term	Description
Term Name	Derived Index
Label	Derived Index
Definition	A derived index is an index that is derived from an existing index
Comment	Derived indices may be the result of structural or functional derivation, or a combination of the two
Type of Term	Derived Index

Term	Description
Term Name	Functional Property
Label	Functional Property
Definition	Functional property refers to the functionality and purpose of the index that is being derived from
Comment	A functional property is different from other properties of an index in that it is in some ways indexes or duplicates, summarizes, supplements, and/or combines existing indices' functionality
Type of Term	Property

Term	Description
Term Name	Structural Property
Label	Structural Property
Definition	Structural property refers to the structural aspect of the index
Comment	A structural property is different from other properties of an index in that it is in some ways mimics the look and feel of the index or indices it is derived from
Type of Term	Property

Term	Description
Term Name	Copy Index
Label	Copy Index
Definition	A copy index is an index that duplicates a portion of an existing index
Comment	Copy index is different from Carbon copy index and from other derived index in that it copies some or all of an existing index's functionality, structure or both
Type of Term	Derived index

Term	Description
Term Name	Carbon Copy Index
Label	Carbon Copy Index
Definition	A carbon copy index is an index that duplicates an existing index a hundred percent
Comment	Carbon Copy index is different from copy index and other derived index in that it is an exact copy of an existing index or indices
Type of Term	Derived index

Term	Description
Term Name	Meta Index
Label	Meta Index
Definition	An index of an index, meaning an index that is abstracted from the original index without revealing identifiable information, that may be doubly distanced from the original base index or other derived index
Comment	Meta index is different from other derived index in that it is meta of derived index. An example of a meta index is a bibliography of books and book chapters in a particular subject area. It is a meta index of indices
Type of Term	Derived index

Term	Description
Term Name	Abridged Index
Label	Abridged Index
Definition	An abridged index is a briefer version of an existing index
Comment	Abridged index is different from other derived index in that it is shorted version of a derived index. An example of a abridged index can be seen in books which poses two indices: one that only lists the chapters and another which lists the chapters along with every section and subsection of each chapter...
Type of Term	Derived index

Term	Description
Label	Fragment Index
Term Name	Fragment Index
Definition	A fragment index is an incomplete portion of an index
Comment	Fragment index is different from other derived index in that it is an incomplete derived index. An example of a fragment index is a table of contents of a book which is missing a page
Type of Term	Derived index

Term	Description
Label	Supplemental Index
Term Name	Supplemental Index
Definition	A supplemental index is an index which supplements an existing index
Comment	An example of a supplemental index in a book is seen in various paper copy textbooks which include several additional (not included on paper) chapters on either an included digital media such as a disk or the textbook's/author's/publisher's website
Type of Term	Derived index

Term	Description
Label	Tangential Index
Term Name	Tangential Index
Definition	A tangential index is an index which indexes related data to data already indexed in an existing index
Comment	For example, a book word index, listing the pages upon which a word occurs, can be enhanced with a tangential index that adds the line number or numbers in that page wherein the word occurs
Type of Term	Derived index

Term	Description
Term Name	Intersectional Index
Label	Intersectional Index
Definition	An intersectional index is an index which combines parts of existing indices into a single index
Comment	For example, a table of contents in a book that is an editorial work which brings together parts of several books and lists several chapters, sections, and subsections of each in one uniform index is an intersectional index
Type of Term	Derived index

Term	Description
Term Name	Union/Aggregate Index
Label	Union/Aggregate Index
Definition	A union/aggregate index is an index which combines 2 or more indices in their entirety into a single index
Comment	Any book series which after completion is then republished in a single volume edition will contain a master index which brings together the indices of each of the books together in a single index
Type of Term	Derived index

References

1. Ahmed, I., Fayyaz, A., Shahzad, A.: PostgreSQL Developer's Guide. Packt Publishing Ltd., Birmingham (2015)
2. Ball, A., Darlington, M.: Briefing paper: the adobe extensible metadata platform (XMP). UKOLN research organization (2007)
3. Choi, K.R., Kim, K.C.: T*-tree: a main memory database index structure for real time applications. In: Proceedings of 3rd International Workshop on Real-Time Computing Systems and Applications, pp. 81–88. IEEE (1996)
4. Dwyer, C.: Privacy in the age of google and facebook. IEEE Technol. Soc. Mag. **30**(3), 58–63 (2011)
5. Foote, K.D.: A brief history of metadata. Online (Data Varsity) (2019)
6. Garcia-Molina, H., Ullman, J.D., Widom, J.: Database System Implementation, vol. 672. Prentice Hall, Upper Saddle River (2000)
7. Gilliland, A.J.: Setting the stage. Introduction Metadata **2**, 1–19 (2008)
8. Gitelman, L.: Raw Data is an Oxymoron. MIT Press, Cambridge (2013)
9. Griffel, D.M., McIntosh, S.D.: Admins: a progress report. Technical report. Center for International Studies, Massachusetts, Cambridge (1967)
10. Heumann, K., Mewes, H.W.: The hashed position tree (HPT): a suffix tree variant for large data sets stored on slow mass storage devices. In: 3rd South American Workshop on String Processing, pp. 101–114 (1996)
11. Hey, T., Trefethen, A.: The data deluge: an e-science perspective. In: Grid Computing: Making the Global Infrastructure a Reality, pp. 809–824 (2003)

12. Japp, R.: The top-compressed suffix tree: a disk-resident index for large sequences. In: Proceedings of the Bioinformatics Workshop at the 21st Annual British National Conference on Databases (2004)
13. Lawrence, B., Lowry, R., Miller, P., Snaith, H., Woolf, A.: Information in environmental data grids. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. **367**(1890), 1003–1014 (2009)
14. Lehman, T.J., Carey, M.J.: A study of index structures for main memory database management systems. Technical report, University of Wisconsin-Madison, Department of Computer Sciences (1985)
15. Orr, D.A., Sanchez, L.: Alexa, did you get that? Determining the evidentiary value of data stored by the Amazon® echo. Digit. Investig. **24**, 72–78 (2018)
16. Phoophakdee, B., Zaki, M.J.: Genome-scale disk-based suffix tree indexing. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 833–844 (2007)
17. Riley, J.: Understanding metadata. National Information Standards Organization, Washington DC, United States, p. 23 (2017). http://www.niso.org/publications/press/UnderstandingMetadata.pdf
18. Şah, M., Wade, V.: Automatic metadata mining from multilingual enterprise content. J. Web Semant. **11**, 41–62 (2012)
19. Schmidt, D.: Google data collection. Digital Content Next [Online] (2018)
20. Smith, J.R., Schirling, P.: Metadata standards roundup. IEEE Multimed. **13**(2), 84–88 (2006)
21. Voigt, P., Von dem Bussche, A.: The EU general data protection regulation (GDPR). A Practical Guide, 1st edn. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-57959-7
22. Wheatley, H.B.: What is an Index?: A Few Notes on Indexes and Indexers, vol. 1. [for the Index Society]. Longmans, Green (1879)