Link to the Balisage Proceedings Home Page at http://www.balisage.net/Proceedings

Link to this paper (entry page with links) at http://www.balisage.net/Proceedings/vol25/html/Huitfeldt01/BalisageVol25-Huitfeldt01.html

BalisageThe Markup Conference

Balisage Paper: Document similarity

Transcription, edit distances, vocabulary overlap, and the metaphysics of documents

Claus Huitfeldt

Associate Professor

University of Bergen

<Claus.Huitfeldt@uib.no>

C. M. Sperberg-McQueen

Founder and principal

Black Mesa Technologies LLC

Balisage: The Markup Conference 2020
July 27 - 31, 2020

Copyright ©2020 by the authors. Used with permission.

How to cite this paper

Huitfeldt, Claus, and C. M. Sperberg-McQueen. “Document similarity: Transcription, edit distances, vocabulary overlap, and the metaphysics of documents.” Presented at Balisage: The Markup Conference 2020, Washington, DC, July 27 - 31, 2020. In *Proceedings of Balisage: The Markup Conference 2020*. Balisage Series on Markup Technologies, vol. 25 (2020). https://doi.org/10.4242/BalisageVol25.Huitfeldt01.

Abstract

In recent years, development of tools and methods for measuring document similarity has become a thriving field in informatics, computer science, and digital humanities.

Historically, questions of document similarity have been (and still are) important or even crucial in a large variety of situations. Typically, similarity is judged by criteria which depend on context.

The move from traditional to digital text technology has not only provided new possibilities for discovery and measurement of document similarity, it has also posed new challenges. Some of these challenges are technical, others conceptual.

This paper argues that a particular, well-established, traditional way of starting with an arbitrary document and constructing a document similar to it, namely transcription, may fruitfully be brought to bear on questions concerning similarity criteria for digital documents. Some simple similarity measures are presented and their application to marked up documents are discussed. We conclude that when documents are encoded in the same vocabulary, n-grams constructed to include markup can be used to recognize structural similarities between documents.

Table of Contents

Introduction

In recent years, development of tools and methods for assessing document similarity has become a thriving field in informatics, computer science, and digital humanities. A plethora of methods has emerged, typically designed to calculate some quantitative measure of the similarity between documents. Such measures may be useful for document version control, authentication, summarization, information retrieval, and so on. Judgments of document similarity have also been (and still are) central in contexts like library cataloguing, accusations of plagiarism (whether in literature, education, or research), assessment of documentary evidence in criminal or civil courts, version control in administrative document archives, authentication of identity documents, historical study, and literary studies.

What is meant by “similarity”, or closely related words like “sameness” or “identity” of documents varies widely from situation to situation. A document which in one context is seen as so similar to another document that it constitutes a case of plagiarism, may in another context be judged as too different to count as a faithful rendering of the same document. It is only natural, therefore, that also the methods or criteria by which we decide whether two documents count as similar should vary from situation to situation.[1]

Given such variations in what counts as similarity, it may be foolish to seek, as we do here, for general principles of document similarity which hold in all contexts; we hope, however, that the results may justify the search. We begin by trying to determine what document similarity *is* (in section “What *is* document similarity?”), and then (in section “Some existing measures of document similarity”) turn to a survey of some existing methods of measuring document similarity quantitatively. There is a noticeable gap between the models of documents assumed by those methods and the models underlying most descriptive markup systems, and we suggest (in section “What to do about marked up documents?”) ways in which the gap might be bridged. With this background in place we can present the results of some simple concrete tests of various measures as applied to different kinds of document similarity.

What *is* document similarity?

Similarity and identity

Similarity comes in degrees. In everyday language, the word "similar" is often related to words like "identical" and "different", for example when we say that two things are "so similar that they are in effect identical", or when we say that they are so dissimilar that they are "completely different".

The similarity measures presented later may inform us that documents A and B have a similarity score of 0.89, while documents A and C are only 0.67 similar, measured on a scale from 0 to 1. As values, both 0 and 1 may put some strain on our imagination. The value 1 would indicate complete similarity, and 0 complete dissimilarity.

It may be hard to think of examples of *complete dissimilarity*: For any two objects,[2] there are always innumerable ways in which they may be similar. Any two randomly picked documents, for example, may be similar regarding language, vocabulary, compositional, linguistic or rhetorical structure, theme, genre, style, plot, narrative, dominant meter, and so on. When it comes to digital documents (or digital objects in general) they will at least be similar in that there is some specific byte or bit value that they do both contain -- or not contain.

When we consider digital objects, it may perhaps seem easier to think of examples of *complete similarity*. For example, two documents may consist of exactly the same sequence of words, or characters, or bytes, or even bits. But even then, they are similar only at a certain level of abstraction. Considered as physical representations of sequences of bits, bytes, characters etc., they will at least be different in that they occupy different portions of some physical data carrier.[3]

We observed that similarity comes in degrees. Identity does not, at least not according to the standard philosophical concept of identity: the only way for two objects to be identical is for them to share all their properties. And if they do, they are not two objects at all, but one and the same object.

As formulated, this assumes that identity is a question of sharing of properties. Furthermore, it assumes not only the so-called indiscernibility of identicals ($x=y \Rightarrow (Fx \Leftrightarrow Fy)$), but also the identity of indiscernibles ($(Fx \Leftrightarrow Fy) \Rightarrow x=y$).[4] If A and B are identical, "A" and "B" are just two names for the same object. So the discovery about the relation between the Evening Star and the Morning Star was not a discovery that two objects have all the same properties, but that they are the same object known under two different names.[5]

Identity is an equivalence relation, i.e. it is reflexive, symmetric and transitive. Unlike other equivalence relations, however, identity is a relation that every object has to itself and to no other object. Some common views are that concrete objects are identical if and only if they occupy the same region of space and time, abstract objects if they share all their predicates, and sets if they have exactly the same members.

Like identity, the concept of similarity also has a prominent place in philosophy, and many or most attempts to explain what similarity is draw on the notion of identity. While identity is similarity in regard to all properties, one might say, similarity is identity with regard to some subset of properties.

Unfortunately, this way of defining identity and similarity leads to certain paradoxes.[6] In particular, certain problems with the notion of similarity made Nelson Goodman, among others, formulate a critique [Goodman] which cast widespread doubt on the usefulness of the concept of similarity as an analytic tool.

A slightly modified concept of similarity (and identity) which arguably avoids these paradoxes, and at any rate is more useful for our purposes, is one according to which similarity is defined as similarity in some respect, where the respect in which some things are judged similar is always decided by context [Douven and Decock 2010]. In this conception of identity and similarity, both relations are intransitive and context-sensitive relations. (In addition to being, according to the authors, vague and “somehow” subjective.)

So, on this view, which is also the view on which our discussion is based, in order to make a meaningful statement that two objects are similar, or that they are similar to a particular extent, it is necessary always to make clear, referring to context, in respect to which property the similarity is supposed to hold.

What is a document?

If similarity is always and only to be understood as similarity with respect to some particular property or set of properties, and if the properties an object can possess depend on the nature of the object (as on most accounts they do), then any measure of document similarity necessarily reflects some underlying model of documents and their properties.

It is common to distinguish documents from texts by saying that a document is a concrete object instantiating an abstract object, the text. This may itself be seen as an account or statement of similarity: the many copies of a book are similar, and may for some purposes even be considered identical, *because* they are all instances of the same abstract object, i.e. the text.

Since abstractions can operate on different levels, i.e. be more or less inclusive, this way of accounting for similarities allows for example bibliographers to distinguish between different items, print-runs, editions, translations etc. of the same work.[7]

The distinction can be (and has been) applied equally well to digital documents, though with a couple of noticeable differences. When the distinction is applied to traditional documents, questions are often raised concerning the status of the text, and how we can interact with abstract objects. When the distinction is applied to digital documents, questions tend to be raised concerning the status of the document side of the equation as well: The notion of a digital document may itself seem to involve abstraction.[8]

In our discussion of document similarity, and also of transcription, we will employ a distinction similar to the one between abstract objects and their instantiations, namely the distinction between types and their tokens.

This distinction was introduced by Peirce, when he observed that even though there is only one word (type) “the” in English, there may be many tokens of that word on any page. [Peirce] If we loosely call any perceivable pattern on a page a “mark”, then some of those marks are tokens, namely those which instantiate types.

Each token instantiates one and only one type within a given repertoire of types, though tokens of one repertoire (e.g. the letters of the alphabet) may combine to form composite tokens of composite types (e.g. English words, sentences or documents). One and the same token may instantiate types within different type repertoires (such as “I” which may be a token both of the Latin letter I, the English first person singular pronoun, and a sentence).

There is no requirement that different tokens of the same type be perceptually similar. For example, type identical tokens of handwritten, typewritten, printed or digital documents may look very different. Nor is it always easy to identify features possessed by all tokens of a type. If there are any properties shared by all instances of lowercase *G*, for example, in all fonts and all standard forms of handwriting, for instance — other than being an instance of lowercase *G* —, they are not immediately obvious.

Two further questions demand brief discussion: what are documents made of? and how are those constituent parts organized to make up the document? We do not seek to settle these questions here, only to identify several different possible answers, each of which leads to a different model of documents, and thus to a different set of possible properties and a different concept of document similarity.

- In ways which we hope will become clear in what follows, different measures of document similarity may model documents as being made up either of tokens or of types; since tokens are concrete objects and types are not, strictly speaking one should perhaps distinguish between token-based measures of document similarity and type-based measures of text similarity.
- Whether taken as types or as tokens, the constituent parts of a document may be identified as the characters, or words (or whitespace-delimited character sequences, with or without removal of punctuation), or nodes of various kinds appearing in the document.[9]
- Existing similarity measures appear to postulate several different organizational principles for documents, and others are familiar from the literature on the theory and practice of markup. Documents may be modeled as sequences of tokens or types, as sets, as bags, as ordered or unordered trees, or as directed graphs or hypergraphs. This is not a complete catalogue, but it may be enough to go on with.

Transcription and t-similarity

A different approach to the question of document similarity might start from purely practical considerations. For centuries, documents were reproduced by transcription, with the implicit requirement that the transcript be similar to the exemplar in whatever respects were necessary for the purpose at hand. Historically, and up until our own time, one of the reasons for transcription has been to provide better access to unique manuscripts and rare books which could otherwise only be consulted by visiting the source library or archive, or by transportation of the document itself. In such cases, the transcript would be more accessible than the original; it might also be easier to read than the exemplar and thus provide access for a wider audience. Transcription was often the first step towards a printed edition of the source, which would thus have been made available to an even wider audience.

Transcription, as it is practised in textual criticism, often as part of the preparation of a documentary or critical edition of some body of documents, is sometimes described as the attempt to reproduce a particular document *as faithfully as possible*.[10]

In scholarly contexts, transcripts were (and still are) important because for some purposes they can serve as substitutes for the exemplar, i.e. the original document. No transcript can preserve all properties of the exemplar. But in order to be usable as a substitute for the exemplar in any given context, the transcript must reproduce with complete accuracy the properties of the exemplar relevant to that context. With respect to any given context, therefore, transcription preserves similarity of documents with respect to some particular property or properties. For this reason, transcription seems well positioned to shed some light on the nature of document similarity with respect to a particular property.

Our earlier work on the logic of transcription has led us to believe that the essential property of transcripts is that as a general rule transcripts do not reorder, correct, amend, or normalize the text of the exemplar in any way. Generalizing or oversimplifying slightly, we might say that they do not add anything to or remove anything from the exemplar. Thought of in this way, transcription should be the archetype of document similarity. If a document, T, is a transcript of another document, E, then for the relevant purposes (and with respect to the relevant properties) T is *as similar as possible* to E.

One might perhaps have thought that digital text technology, with the availability of cheap, high quality facsimiles, would have made transcription obsolete. What can be a more faithful reproduction, and thus more similar, to the exemplar than a high-quality image? Indeed, digital facsimiles are of tremendous value to textual scholarship, and in many cases they can reduce the need for transcription.

However, a transcript is usually not primarily a reproduction of the visual appearance of a document, but of its textual content. Manuscripts with difficult handwriting and archaic or idiosyncratic spelling may be hard to read for non-specialists. A digital image cannot be analysed and searched in terms of linguistic criteria, but a transcription can. So one might say that digital facsimiles provide visual similarity, but do not necessarily provide the kinds of accessibility and processability we may desire. Achieving a form of similarity compatible with ease of reading, searching, and textual manipulation is a prime concern of transcription.[11]

Optical character recognition comes closer to the aims of transcription. Today most printed and many hand-written documents can be OCR-read with considerable degree of accuracy, and many projects use OCR as part of the preparation of transcripts. It is hard to predict how close OCR-readers may come to human readers on all kinds of documents. As things stand today, they can at least often serve as efficient tools in preparing a transcription.

Transcription of difficult source material requires exceptional reading skills so far found only in human experts. And all transcription requires knowledge of the linguistic, historical and cultural context of the document in question. When visual evidence is inconclusive, as it often is, disagreements about what is the correct transcription of a particularly difficult passage in a manuscript are settled by arguments as to what is the most probable interpretation of the text in question.

We said above that a transcript T of an exemplar E is similar to E, in some very strict sense of similarity. A more precise and formal account of this relationship, which we may call t-similarity, can be given very briefly, in terms of the type-token distinction already introduced:

If T is a transcript of E, then T and E are tokens of the same type.[12]

T-similarity is not designed to capture every aspect of the relation between T and E, but only the specific kind of similarity that must hold between them. Whereas transcription is an irreflexive, asymmetric, and intransitive relation, t-similarity is (under appropriate assumptions) an equivalence relation, i.e. it is reflexive, symmetric and transitive.[13]

The brief, general definition of t-similarity just given conceals considerable underlying complexity. In all but trivial cases, a transcript instantiates a composite type, which may contain tokens of e.g. letters, words, sentences, quotations, cross-references, notes, and comments, some of which are sequentially and/or hierarchically ordered, others only partially ordered etc.

A somewhat more detailed formulation of t-similarity may be given as follows: If T and E are t-similar, then, with certain well-defined exceptions, the following rules apply:

- *Reciprocity:* There is a one-to-one relation between the tokens in an exemplar E and tokens in its transcript T.
- *Purity:* Every token in T transcribes something in E.
- *Completeness:* Every token in E is transcribed by something in T.
- *Type similarity:* corresponding tokens in E and T are tokens of the same type.

We said earlier that typically, transcriptions do not reorder, correct, amend, or normalize the text of the exemplar in any way. The four rules just stated are a way of making this statement precise. Note that if transcription is taken (as it often is) to be the task of creating an artifact which is *as similar as possible* to its exemplar (at least, for the kinds of purposes for which transcription serves), and if these four rules capture the essential properties of transcription, then it follows that these four properties of reciprocity, purity, completeness, and type identity provide an explicit account of what it means for one document to be *as similar as possible* to another (again, within the limits imposed by the context within which transcription is relevant).

If the description of transcription just offered were entirely true, without qualification, we would need to say no more about transcription. In reality, of course, as could perhaps be expected, it is not true without qualifications.

Transcripts usually include material not found in the exemplar, such as page numbers and explanatory notes. Some transcripts omit material found in the exemplar: deletions, marginal remarks or insertions in a different hand. Different transcribers apply different criteria for type identity: some normalize spelling or expand abbreviations, some preserve allographic variation where others do not, while yet others preserve type identity not on letters but only on the word level. Sometimes such deviations from t-similarity are marked, often not.

Differences between transcripts may mean that the transcribers disagree about the content of the exemplar: when one transcript of the manuscripts of the nineteenth-century German writer Ludwig Büchner renders a word as *Woyzeck* and another as *Wozzeck*, it signals a disagreement about how to read Büchner's handwriting.[14] However, differences between transcripts may also simply indicate that the two transcripts follow different transcription practices. If one transcript shows an abbreviation where the other shows a word fully spelled out, it does not mean the two transcribers disagree about what is on the page, only that they have reached different conclusions about how best to make the transcript useful. (Are we trying to support simple textual searches? Abbreviations may need to be expanded. Are we trying to make it possible for students of paleography to find occurrences of particular short forms? Abbreviations may need to be carefully recorded.) The presence, in cultural practice, of different transcription practices does not mean that there are no facts of the matter, or that transcripts cannot agree or disagree about those facts, or that transcripts cannot simply be right or wrong. But it does mean that understanding the logical implications of a given transcript requires a hermeneutic attempt to understand the transcription's practice.

Often transcripts come with descriptions of their conventions in the form of legends or statements of practice. Typically such statements describe deviations from what is assumed to be general norms of practice, or norms within an implied community of practitioners. Those general norms themselves are usually not stated explicitly, perhaps because they seem to be too obvious to allow mention without offence to the reader. What such statements of practice seem to describe, in addition to deviations from norms of the community, are exceptions from our account of t-similarity.

We believe, that is, that the three properties of t-similarity (purity, completeness, type identity) constitute a sort of "default" transcriptional implicature, a core of tacit assumptions underlying all transcription. Furthermore, we believe that deviations from this default implicature can usefully be described in a systematic way simply by making explicit a distinction often implicitly made, between "normal" and "special" tokens in transcript and exemplar.

Special transcript tokens are tokens which transcribe no token in the exemplar. Special exemplar tokens are tokens which are not transcribed by any token in the transcript. With suitably formulated rules for the interpretation of special transcript and exemplar tokens, a fully operationalized account of the t-similarity between transcript and exemplar can be given.

This allows us to judge, in the case of differences between transcripts of the same exemplar, whether they differ only because of differences in transcriptional policy, or because one makes statements on matters about which the other is silent, or whether they actually make different claims about the exemplar.

In what follows, where we present some formal methods for measuring similarity, we will argue that our theory of t-similarity may provide a rationale, or a common theoretical underpinning, for the various methods discussed. They may all be understood to measure, in some way or other, similarity between any pair of documents in terms of their t-similarity, typically by counting how many additions, omissions and substitutions are required to account for the difference between them. The different methods vary, however, in their underlying document model: they model documents as simple or composite tokens or of various types, organized as e.g. sequences, bags, sets or graphs.

Some existing measures of document similarity

This section describes some important measures of similarity which can be and often are applied to problems of document similarity. The survey is by no means complete: we pass in silence over the large topic of statistical methods developed in the course of decades of research in information retrieval and also over the machine-learning methods which currently enjoy wide popularity; neither the time at our disposal nor the current state of our knowledge allows us to address them.

Instead, we limit ourselves to some conceptually simple measures, for which it is possible at least in principle to identify clearly the properties with respect to which document similarity is being measured.

The theory underlying these approaches has been a topic of discussion among mathematicians, computer scientists, and students of information theory for the last seventy-five years or so, and there is a huge body of scholarly work, bristling with mathematical formulae, which you will be relieved to hear we have no intention of attempting to summarize here. Because it makes slightly better sense mathematically to view things this way, rather than talking about document *similarity*, they talk about document *dissimilarity*, or (as it is normally called) *distance*. All the measures discussed in the following are carefully defined to serve as what mathematicians call metrics, or distance measures. So they have several properties worth bearing in mind, and easily understood by analogy with distance in the everyday world.[15]

- For any points A and B, the distance between A and B is a non-negative number.
- For any points A and B, the distance between A and B is zero if, and only if, A and B are the same point.

  This is the mathematical realization of the identity of indiscernibles, which we mentioned above.
- For any points A and B, the distance between A and B is the same as the distance between B and A.

  That is, distance measures are symmetric. There are real-life cases where this is not true: in the Alps, the walking distance from village A to village B is not the same as the walking distance from B to A. In a city with many one-way streets, distance can also be non-symmetric. Quantitative measures for such asymmetric distances sometimes carry the endearing name *quasimetrics*.
- For any three points A, B, and C, the distance between A and B is less than or equal to the sum of the distances from A to B and B to C.

  This is called the *triangle inequality*, because it guarantees that we can draw a triangle in the Euclidean plane with the same relative distances among the points. The triangle inequality is important for certain kinds of reasoning based on distance measures, and notable because not every intuitively plausible idea for a distance measure turns out to obey it.

As mentioned earlier, similarity measures are usually given in terms of a decimal number between 0 and 1. The higher the number, the greater the similarity. Some measures (e.g. the Levenshtein distance we will discuss below) are originally defined not to give values between 0 and 1, but as integer values counting a number of operations required to convert one sequence or set of a specific length to another. In such cases, we may “normalize” the measured value by dividing it by the length of the longest sequence. So if the distance measure between two sequences of lengths 10 is 3, we can normalize the value by dividing it by the length of the longer sequence, i.e. 0.3. Furthermore, we can convert a normalized distance measure to a *normalized similarity measure* by subtracting its value from 1, so e.g. a normalized distance of 0.3 corresponds to a normalized similarity of 0.7.

In the following, all measures will, unless otherwise stated, be given in the form of normalized similarity measures, irrespectively of how they are originally defined.

Sequence-based similarity measures

What sequence-based methods for identifying or measuring similarity between documents have in common is, as the term indicates, that they operate on representations of documents as *sequences* of types, usually sequences of letters or words. The Hamming distance is one of the first sequenced-based similarity measures implemented digitally, while file comparison tools like diff are probably among the most widely used sequence-based methods. We will not go into detail about these here.[16]

An extremely influential distance measure was investigated in 1965 by the Russian scholar Vladimir Levenshtein and is known for him as the *Levenshtein distance*. The Levenshtein distance is the minimal number of primitive operations needed in order to transform one string into another. For this reason, it and similar distance measures are sometimes referred to as *edit distance* measures. (As with all string distance measures, we can apply the Levenshtein distance equally well to sequences of words, or sequences of any other kind of item, but it simplifies the discussion and shortens the examples to talk primarily about sequences of characters.) In the Levenshtein distance, the primitive operations allowed are threefold:

- insertion of a character
- deletion of a character
- replacement of a character by another

The algorithm for calculating the Levenshtein distance is slightly difficult to follow, and is rather expensive (which is why practical file comparison programs typically do not try to apply it in its pure form), but some examples should serve to illustrate the measure; we include the Hamming distance, to ease comparisons.[17]

Table I

Strings	Levenshtein distance	Hamming distance	Levenshtein similarity	Hamming similarity
plagiarism, plagiarsim	2	2	0.80	0.80
nationals, anational	2	9	0.78	0.00
Ruhrgebiet, Ruhegebete	3	4	0.70	0.60
Merkel, Markle	3	3	0.50	0.50

As may be seen, the Levenshtein distance is lower than the Hamming distance for the second and third pairs — and for the second pair, much lower.

Few discussions of edit distance spend much time on any justification for the choice of primitive operations, but some reflection should make clear that the three operations chosen suffice for changing any string into any other string.

More than that, they correspond directly to three of the properties identified above as characteristic of error-free transcriptions and thus of maximally similar documents: purity is the absence of insertions; completeness is the absence of deletions; type identity is the absence of character replacements. We suggest that this explains why the Levenshtein distance is universally and intuitively accepted as a measure of similarity.

Two strings of tokens are maximally similar and have a Levenshtein distance of zero just when they instantiate exactly the same sequence of types. In a sense, the Levenshtein distance operates by counting the number of violations in the two strings of the properties of purity, completeness, and type identity.[18]

A suggestive variant on the Levenshtein distance is the Damerau/Levenshtein distance, named for Fred J. Damerau (and Vladimir Levenshtein), who in 1964 published a study of spelling correction in which he wrote that 80% of the typographic errors in the materials he studied consisted of a single inserted letter, a single dropped letter, a single changed letter, or a single transposition of two adjacent letters.[19]

We find the Damerau/Levenshtein distance suggestive in part because its difference from the Levenshtein distance lies in a tacit (and partial) recognition that there are two distinct levels of object to consider here: the identity of the sequence, and the identity of its constituent characters. This seems to us plausible and suggestive for considerations of document similarity as distinct from string or sequence similarity. (We distinguish document similarity from string similarity because we do not believe that documents can be satisfactorily viewed as flat unidimensional sequences of items — whether the items are paragraphs, sentences, words, or characters). As mentioned above, our work on transcription similarity has found it helpful to consider not a single level of types and tokens, but many levels of composite types and tokens, containing and organizing lower-level types and tokens, down to a level at which analysis stops.

Once we explicitly recognize that multiple levels of type and token are involved in documents, we can easily understand some alternative similarity measures as selectively ignoring certain types and tokens and basing the measure on others. Other similarity measures involve variations on the concept of type identity (or type similarity), often by assigning different weights to different types and tokens.

Set- and bag-based similarity measures

Perhaps the simplest alternatives to the sequence measures just discussed are measures that ignore the fact that the items in question (characters, or in many application word types) form a sequence at all. If we pay attention only to which types appear in the sequence, ignoring the sequence in which they appear, then we can apply measures designed to quantify the similarity between sets.

Like Levenshtein distance, the measures discussed in this section can be thought of as counting exceptions to t-similarity (i.e. additions, omissions and substitutions). The difference is that the Levenshtein distance counts changes to sequences, and these measures count changes to sets or bags of types or tokens.

A widely used measure for set similarity is *Jaccard similarity*.[20] It counts first the number of items that appear in both sets (the cardinality of their intersection), and then the number of items that appear in either or both (the cardinality of their union), and then divides the one by the other. In a formula: $|A \cap B| / |A \cup B|$.[21]

> An obvious refinement to this approach is to keep track not just of which character types appear in the two strings, but how often they are used: that is, to treat the string as a set of character *occurrences*, or as is more commonly said, a *bag of characters*.[22] This may allow us to register the difference between strings in which a particular character occurs frequently from documents in which it is used only once.[23]

Applied to our examples, the Jaccard set and bag measures produce the following results:[24]

Table II

Strings	Jaccard set	Jaccard bag	Levenshtein	Hamming
plagiarism, plagiarsim	1.00	1.00	0.80	0.80
nationals, anational	0.86	0.80	0.78	0.00
Ruhrgebiet, Ruhegebete	0.78	0.67	0.70	0.60
Merkel, Markle	0.83	0.71	0.50	0.50
commas, come	0.50	0.43	0.50	0.50
actresses, recast	1.00	0.67	0.22	0.00
thread, hair	0.43	0.43	0.17	0.00

As was perhaps to be expected, Levenshtein similarity is more sensitive to the order of the characters in the strings than either of the Jaccard measures, but the Jaccard bag measure deviates less from Levenshtein than the set measure. It can also be observed that, with a small exception for the third and a big exception for the sixth pair, the Jaccard measures give the same relative ranking of similarities as the Levenshtein measure.

In addition to the Jaccard similarity, many other similarity measures can be applied to sets and bags. We list here all the ones we have considered in this work. In the following formulas, *A* and *B* are the two collections being compared, |*A*| is the cardinality of *A*, and min(X, Y) is the smaller of X and Y.

- Jaccard similarity: $|A \cap B| / |A \cup B|$ (size of intersection over size of union). May in summaries be labeled Js (for sets) or Jb (for bags).
- Symmetric subsumption similarity[25]: $(|A \cap B| / |A| + |A \cap B| / |B|) / 2$ (average of size of intersection over size of the two collections). Labeled as Hs, Hb.
- Szymkiewicz-Simpson similarity[26]: $|A \cap B| / \min(|A|, |B|)$ (size of intersection over size of smaller collection). Labeled as Os, Ob.
- Dice/Sørensen similarity[27]: $2 * |A \cap B| / (|A| + |B|)$ (twice the size of the intersection, over the sum of the sizes of the two collections). Labeled Ds, Db.
- Cosine similarity[28]: $|A \cap B| / \mathrm{sqrt}(|A| * |B|)$ (interprets each set or bag as a vector and calculates the cosine of their angle in an *n*-dimensional Euclidean space). Labeled Cs, Cb.

In the interest of completeness, we provide an overview of the values of these various similarity measures applied to the earlier examples of pairs of strings here (for comparison, the column gives Levenshtein similarity values):

Table III

Strings	Js	Hs	Os	Ds	Cs	Jb	Hb	Ob	Db	Cb	L
plagiarism, plagiarsim	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80
nationals, anational	0.86	0.93	0.93	0.92	0.93	0.80	0.80	0.89	0.89	0.89	0.78
Ruhrgebiet, Ruhegebete	0.78	0.89	1.00	0.88	0.88	0.67	0.80	0.80	0.80	0.80	0.70
Merkel, Markle	0.83	0.92	1.00	0.91	0.91	0.71	0.83	0.83	0.83	0.83	0.50
commas, come	0.50	0.68	0.75	0.67	0.67	0.43	0.63	0.75	0.60	0.61	0.50
actresses, recast	1.00	1.00	1.00	1.00	1.00	0.67	0.83	1.00	0.80	0.82	0.22
thread, hair	0.43	0.63	0.75	0.60	0.61	0.43	0.63	0.75	0.60	0.61	0.17

Until now, we have been dealing with examples of similarity measures applied to strings analyzed as sequences, sets or bags of character types. It should be observed, therefore, that the various similarity measures may behave differently when applied to strings analyzed as sequences, sets, or bags of strings or word types.

Calculating sequence-based measures like the Levenshtein similarity requires time and space proportional to the product of the lengths of the two sequences. For documents of normal length, that can make sequence-based measures prohibitively slow. From a practical point of view, set- and bag-based measures are attractive because they are relatively straightforward to calculate. Perhaps surprisingly, however, simple measures of this kind can and do produce useful results, as will be illustrated below.

In any sufficiently long documents written in an alphabetic script, it is likely that every letter of the alphabet will at some point make an appearance, so that any two documents will have a very high set similarity measure on characters. And any pair of naturally occurring documents in the same language will normally share many high-frequency words, so that no two documents are likely to have a very low bag similarity measure on characters.

But if we take word types, not letter types, as the members of the two sets, we get a comparison of the vocabulary of the two documents, of the kind that has long been used by literary historians to compare and contrast authors or works and even to trace literary influence.

On the principle that a book about a given topic will use terms related to that topic frequently, and terms related to tangential topics only rarely, the idea of comparing documents using the bag of words model was early examined by developers of information retrieval systems.

But since the meaning of a document depends critically upon the meanings of its sentences, and the meaning of a sentence depends, in many languages (notably English, but not only in English), upon the order of its words, it may seem unlikely that a similarity measure which systematically throws away information about the order of words has any chance of producing useful results.

Sequence information without sequences: n-gram similarity measures

One way to make set- and bag-based measures at least partly sensitive to the sequence in which items appear is to model a document not as a set or bag of words or characters but as a set or bag of word or character n-grams.

To illustrate this approach, let us consider the sequence *abcdefghij* and three other sequences derived from it:

- *axbcdefghi*, obtained by inserting an *x* and deleting the *j*.

fghijabcde, obtained by moving the last five characters of the sequence to the front.

jciafgdhbe, a random permutation of the original sequence.

The Levenshtein and Jaccard measures for the three pairs are:

Table IV

Strings	Js	L
abcdefghij, *axbcdefghi*	0.82	0.80
abcdefghij, *fghijabcde*	1.00	0.00
abcdefghij, *jciafgdhbe*	1.00	0.30

As may be seen, the Jaccard measure is unaffected by the permutations of the letters, and the Levenshtein measure is unaffected by ways in which the variants preserve portions of the original sequence.

If we consider not the set of letter types but the set of letter pairs in the strings, however, or the set of letter triples, or subsequences of length n (n-grams), we can get rather different results for these sequences:

Table V

Strings	Js (1-grams)	Js (2-grams only)	Js (3-grams only)	Js (1- and 2-grams)	Js (1-, 2- and 3-grams)	L
abcdefghij, axbcdefghi	0.82	0.78	0.71	0.80	0.78	0.80
abcdefghij, fghijabcde	1.00	0.80	0.60	0.78	0.80	0.00
abcdefghij, jciafgdhbe	1.00	0.06	0.00	0.40	0.26	0.30

As may be seen, including all the n-grams of length n or shorter will produce a higher similarity score than excluding the shorter n-grams. If there are any principled reasons for choosing a particular value for n, we do not know what they are; of course, larger values of n also lead to larger sets and slower computations.

What to do about marked up documents?

The preceding discussion has discussed a number of useful measures for the similarity of sequences, and sets, and bags (and has left many others unmentioned). All of these may be applied to the problem of document similarity by modeling documents as sequences, or sets, or bags, of one kind of thing or another: characters (quite frequent for application of edit distance to words and strings), words, word stems, or units of semantic content created by abstracting away from the words present in the document, using information about word distributions in larger corpora.

Strikingly, painfully absent from this list are measures of document similarity based on, or even passably compatible with, any conception of documents that models them as more complex structures. It has been some time since users of descriptive markup were able to take seriously the idea that documents are in general best modeled as one-dimensional sequences of words or characters. If we translate documents down into such less expressive models for purposes of similarity measurements, we are, it would seem, giving up any chance of using the markup in our documents to guide the measurement of similarity, any chance of measuring the similarity or difference of document structure, to the extent that that structure is expressed by markup.

There has almost surely been research on measuring the similarity of structured documents; we cannot believe there has not been. But we have not found it. Web searches turn up lots of hits for searches like “graph similarity”, but we have not found any which are concerned with similarity of the graphs we use to represent documents, whether the directed graphs of XDM or the directed graphs of GODDAG or the hypergraphs of TAG.

We would have liked to continue our survey of methods for calculating document similarity by considering methods which attend to tree and graph structures at least as carefully as the edit-distance measures for strings attend to the sequence of characters. But we cannot do so, until we either find work on this topic that we have not yet found, or until such measures are invented. Perhaps the best we can do is to identify what seem to us some possible lines of approach which might allow some progress towards document similarity measures which are sensitive to document structure. The following paragraphs attempt to sketch out those lines of approach.

As a short example, consider William Blake's poem *The sick rose*:

> O Rose thou art sick.
> The invisible worm
> That flies in the night
> In the howling storm:
>
> Has found out thy bed
> Of crimson joy:
> And his dark secret love
> Does thy life destroy.

If we wish to measure the similarity of this text to some other text by calculating the Levenshtein distance, or any other edit distance, between the two, we can reduce Blake's poem to a sequence of 172 characters, or to a sequence of 34 words.[29]

1) As a first line of approach, perhaps a more markup-aware similarity measure can be had by simply including the start- and end-tags of elements in the sequence. Depending on our vocabulary and our markup practice, that might produce a sequence of 60 tokens (34 words, 11 start-tags for poem, stanza, and line, 11 end-tags), or more (start- and end-tags for word elements, for example). Perhaps we should insert attribute-value pairs separately into the sequence (as in the line-oriented output from the old SGML parser sgmls). Perhaps we should treat each sequence of data characters (each text node, to shift into XDM terminology) as an item in the sequence, so that (in a simple case with no line-internal markup) each line in the poem becomes a token.[30]

2) A second line of approach is similar: in applying set- and bag-based measures of similarity, we need not restrict ourselves to words or characters. We know we can model Blake's poem as a bag of words. We can equally well model it as a bag of nodes; text nodes can be used whole or broken up into words,[31] which in the case of Blake's rose might give us a bag with 19 members (eight text nodes, one for each line, eight *line* elements, two *stanza* elements, one *poem* element), or one with 45 members (34 words, 11 element nodes) — or perhaps a different count. Whether it is more useful to constitute the bag or set one way or a different way is, of course, a practical question to be settled by experiment.

3) A third line of approach takes the previous idea one step further. An undirected graph is a pair of sets: one set is the vertices of the graph, and the other is a set of pairs whose members are vertices. For directed graphs, one set is the set of nodes, and the other set is a set of ordered pairs whose members are nodes. Alternatively, we can in each case speak of one set of vertices or nodes, and a relation on them.

If we are seeking a way to compare the similarity of document structures we have modeled as graphs, and graphs are simply pairs of sets, then perhaps all we need to do is apply existing measures for similarity of sets to the sets which make up the graphs. (And maybe all that work on graph similarity is relevant to documents, after all.)[32] This is, mutatis mutandis, analogous to the grouping of character and word tokens into types. For our experiments, we have often treated identity of the element type name [or: generic identifier] as a necessary and sufficient condition for node equivalence, but other criteria might be used; we have also experimented with treating all XDM nodes of the same type as equivalent (so any two element nodes are equivalent, any two attribute nodes equally so, and so on).

Of course, the graphs used in document modeling tend to impose orderings on their nodes: the element structure of XML defines an ordered tree, not an unordered one, and similar observations apply to Goddags and to TAG models. But it has been observed that the full XPath data model can be applied to any set of objects to which we can apply two primitive relations which satisfy a short list of conditions; the simplest case to imagine is that one relation defines the *first-child* relation of XDM, and the other a relation one might call *next-sibling*. All the axes of XPath can be defined on the basis of these two primitive relations.[33]

As we observed in our discussion of t-similarity, our model of transcription assumes the existence of composite tokens instantiating composite types; that model of simple and composition types and tokens is explicitly motivated by considerations of document structure [Huitfeldt, Marcoux and Sperberg-McQueen] If as suggested earlier one way to interpret Levenshtein distance is to say that it counts the number of violations of the rules of purity, completeness, and type identity between corresponding tokens within a sequence, then perhaps we can go beyond the sequence by extending the principle to all types and tokens in the documents being compared: how many tokens at any level lack a corresponding item in the other document? How many corresponding tokens differ in type? Of course, finding the minimal number of violations to count is largely dependent on the quality of the correspondence established between the tokens of the two documents; different correspondences need to be explored. A similar problem applies to sequence, and the algorithm for calculating Levenshtein distance can be seen as exploring all possible alignments of the two strings. Perhaps an analogous technique can be applied here. Or perhaps the greater structural complexity of graphs will make it harder to do so without a combinatorial explosion.

4) Yet another line of approach is to step back and think about the relation of Levenshtein distance to sequences. Levenshtein distance measures a distance between two sequences of objects, and it does so very effectively. But neither in the high-level description of what the distance measures nor in the low-level definition of the primitive operations are sequences explicitly mentioned. Could we define a structure-sensitive measure of document similarity analogously, by specifying a set of minimal graph-editing operations and defining the distance measure as the minimal number of editing operations needed to transform one document into the other?

That is, perhaps, a silly question. Of course we can. What is not clear to the authors at the moment is whether one can do so in such a way as to produce a measure of document similarity useful in practice. That involves selecting primitive operations that make intuitive sense and produce distance measures that correspond at least roughly to our intuitions, and it involves defining the measure in detail in such a way as to make it computable, preferably at an acceptable cost.

Further work is needed in this area. In the experiments reported below, we have begun to explore, although only to a very modest extent, some of the lines of approach sketched above.

Some simple experiments

It may be informative to examine the performance of various measures on simple tasks. We have created several test beds, small collections of documents with certain known relations among them. We list the collections here, since some of them will be referred to more than once in what follows.

1. *Testbed 1* consists of 12 articles downloaded from Wikipedia, whereof two sets of three articles and one set of two articles are earlier and later versions of the same document.[34]

   The articles are further subdivided into 312 paragraphs and 1300 sentences.
2. *Testbed 2* consists of 16 paragraphs from the Wittgenstein Nachlass, all rather short.[35]
3. *Testbed 3* consists of 7 articles from the Balisage Proceedings, none of them genetically related. Three of the articles are classified by the Balisage website to be related to the same topic, “DITA”.[36]
4. *Testbed 4* consist of 31 short songs, quatrains, clerihews, and limericks encoded in TEI; some examples involve overlapping hierarchies and are encoded using a variety of markup idioms such as Trojan horse, fragmentation with virtual elements etc.

Identifying genetically related documents

A simple use for document similarity measures, and one of those which first raised our interest in the topic, is that of finding genetically related texts or portions of texts.

Wittgenstein's Nachlass consists of roughly three million word tokens distributed over roughly 54,000 paragraphs. Many of the paragraphs are earlier versions or later revisions of other paragraphs. It is reasonable to expect that these paragraphs are similar in some particular respect, and it is a matter of some practical interest to find a similarity measure which will assign a high similarity score to pairs of paragraphs which are genetically related in this way, and lower similarity scores to paragraphs unrelated by origin or topic.

A number of methods have been employed in order to identify such pairs; the application of the bag of words model is one of the simplest of these. It has turned out to give surprisingly interesting results.

Consider the following three short paragraphs from Wittgenstein's Nachlass:

1. Der Begriff „und so weiter‟ ist äquivalent mit dem Begriffe der Operation.
2. Der Begriff der Operation ist äquivalent mit dem Begriff „und so weiter‟.
3. Der Satz ist der Ausdruck der Übereinstimmung und nicht Übereinstimmung mit den Wahrheitsmöglichkeiten der Elementarsätze.

We expect any reader will note (even if they do not read German) that the first two paragraphs are similar, and the third is not very similar to either. And although they work on such a simple level, the Jaccard similarity measures are effective here: in the bag-of-words and set-of-words models, the Jaccard similarities are as given in the following table. Levenshtein distances measured on characters and on words (after stripping punctuation) are given for comparison.

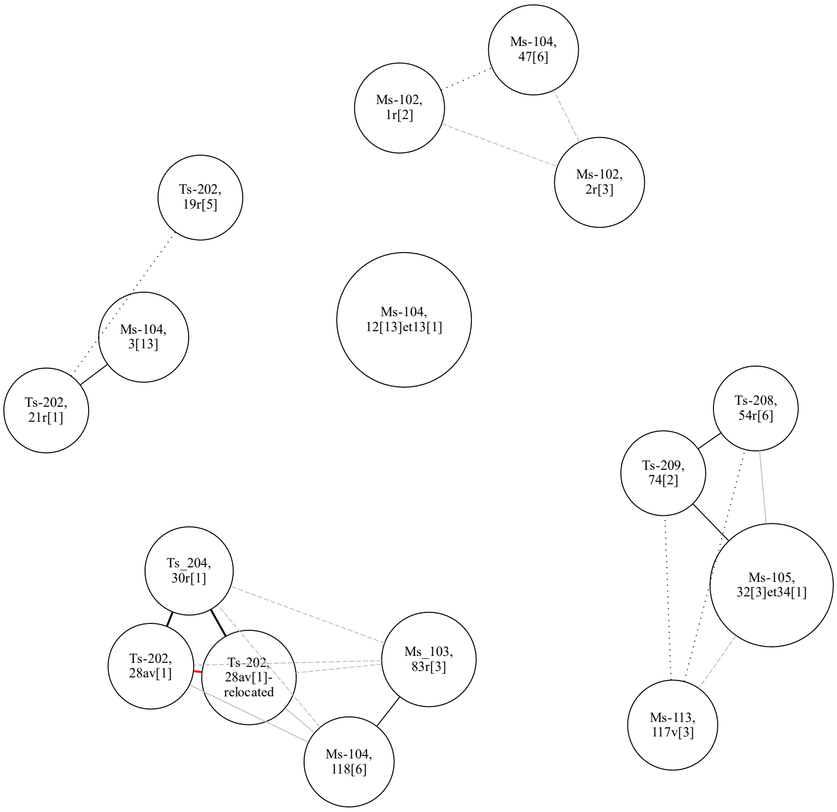
Table VI

Sentences	Js (words)	Jb (words)	L (words)	L (characters)
1, 2	0.92	0.85	0.42	0.64
1, 3	0.26	0.23	0.20	0.32
2, 3	0.28	0.23	0.13	0.33

And indeed paragraph 2 is a re-working in Manuscript 104 (the so-called ‘Proto-Tractatus’ of paragraph 1, which appears in Manuscript 103; paragraph 3 is also from Manuscript 104, but occurs in a different context and is not directly related to paragraphs 1 and 2.[37]

Calculating pairwise similarity for several paragraphs can produce useful visualizations, which (in our testing so far) produce plausible clusterings. If we plot each paragraph of testbed 2 as a point and place them on the Cartesian plane so that the similarity of any two paragraphs is at least roughly reflected in their distance, we get a display of their Jaccard similarity on sets (Js) like the following:

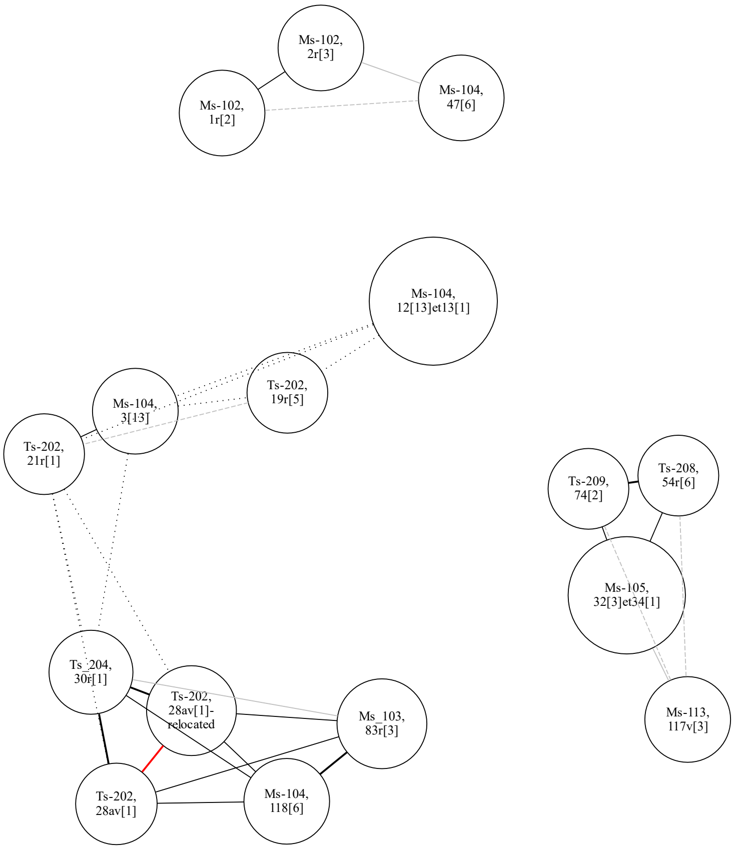
Figure 1: 16 paragraphs from Wittgenstein Nachlass, Jaccard measure on sets



Here, links connect documents with highest similarity (black for higher, gray for lower scores).[38]

The Symmetric subsumption measure on bags (Hb) produces a similar clustering, as shown in the following diagram:[39]

Figure 2: 16 paragraphs from Wittgenstein Nachlass, Symmetric subsumption measure on bags



The tests done so far indicate that indeed, the higher the score, the more likely that the paragraphs are revisions or earlier versions of each other. For testbed 2, all set- and bag-based measures, whether calculated on single words, bigrams or trigrams, give roughly the same results.

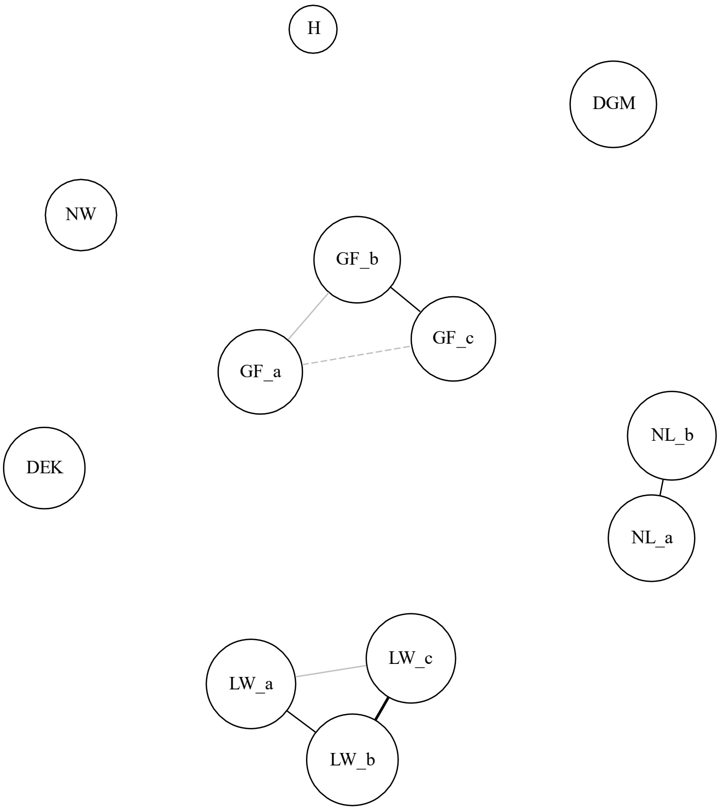
As may be seen, the set and bag method has turned out to have interesting results. How can that be?[40]

An important part of the explanation is that we are comparing relatively short texts. The mean size of paragraphs in the Wittgenstein Nachlass is 58 tokens, and the mean number of types is 41. For any given set of 58 word tokens distributed over 41 word types, the number of ways in which they can be sequenced is astronomic, but the number of ways they can be sequenced into a meaningful sequence of words is severely limited.

But even on longer documents, the bag of words and set of words models are effective in identifying genetically related texts. In fact, somewhat to our surprise, also when applied to testbed 1 (the Wikipedia articles), almost all of the set- and bag-based similarity measures identify the same seven document pairs as most similar to each other, namely those seven document pairs which were genetically related.

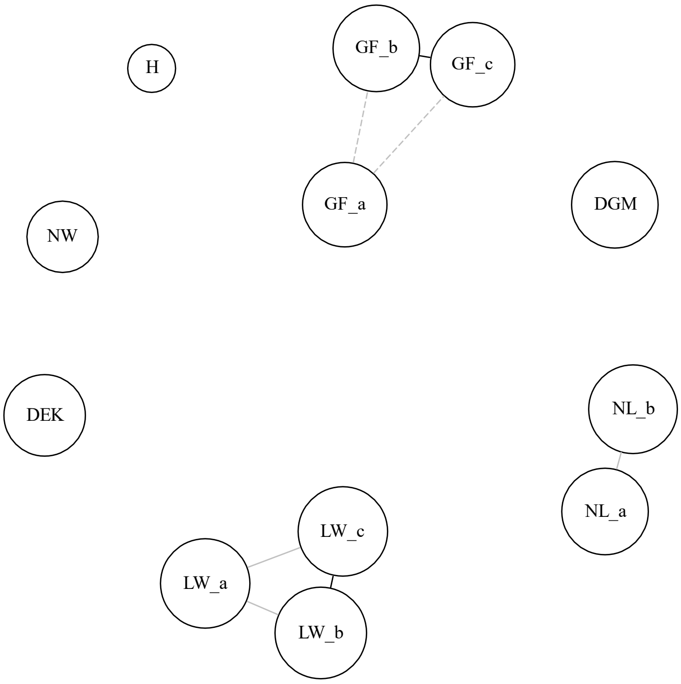
Here, for example, is a plot of the Dice similarity measure on bags. As may be seen, the strongest similarities detected are those of genetically related documents: versions a, b, and c of the article on Gottlob Frege, versions a and b of the article on Niklas Luhmann, and versions a, b, and c of the article on Ludwig Wittgenstein. The articles on Niklaus Wirth, Donald Knuth, and other topics are all isolated.

Figure 3: 12 Wikpedia articles, Dice measure on bags



The Cosine similarity measure calculated on sets of bigrams produces essentially the same result:

Figure 4: 12 Wikpedia articles, Cosine measure on sets of bigrams

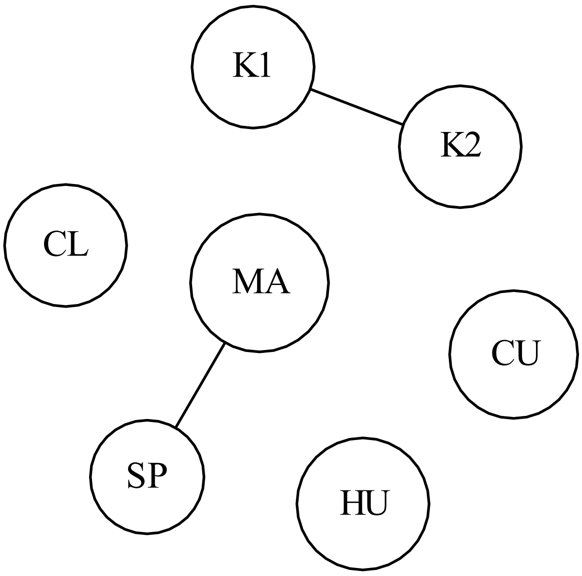


Identifying thematically related documents

Often document similarity is used for information retrieval: given an article on a particular topic and a request to find other articles on the same topic, similarity measures can be used to seek those articles, without any digressions through the difficult thickets of controlled subject vocabularies and the like, which is helpful since indexers and makers of controlled vocabularies find it difficult to keep up with the pace of publication in science and scholarship.

On testbed 3, our sample of 7 documents from the Balisage proceedings series, we ran a variety of tests to see if any would put the three papers indexed as being about DITA together. The results were not encouraging. On the assumption that inclusion of punctuation and case sensitivity and the most frequent word forms, while possibly important for identifying genetic similarity, might have the opposite effect for thematic similarity searches, we omitted punctuation, used cased-insensitive indexing and a stop-list containing all word forms occurring in 6 or 7 of the 7 documents. The results were still not very convincing. Here, for example, is the plot of the Jaccard similarity on bags:

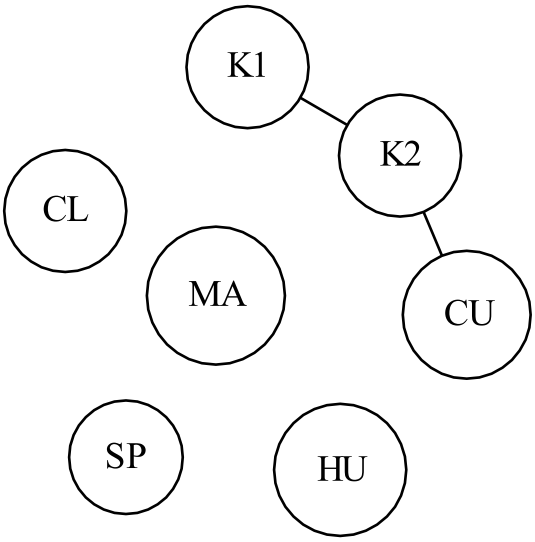
Figure 5: 7 Balisage articles, Jaccard measure on bags



As may be seen, K1 and K2 are indicated as more similar to each other than to the other documents. But their assumed similarity to CU (the “DITA”-papers are CU, K1, and K2) is not reflected, and none of the documents are noticeably closer to each other than to the others.

Several measures identified the similarity between K1 and K2, but only two measures (specifically the Szymkiewicz-Simpson similarity on sets and bags) did assign the highest similarity measure to the two pairs K1-K2 and K2-CU, thus placing the three DITA documents in a cluster, as shown here for the Szymkiewicz-Simpson similarity on bags:

Figure 6: 7 Balisage articles, Szymkiewicz-Simpson measure on bags



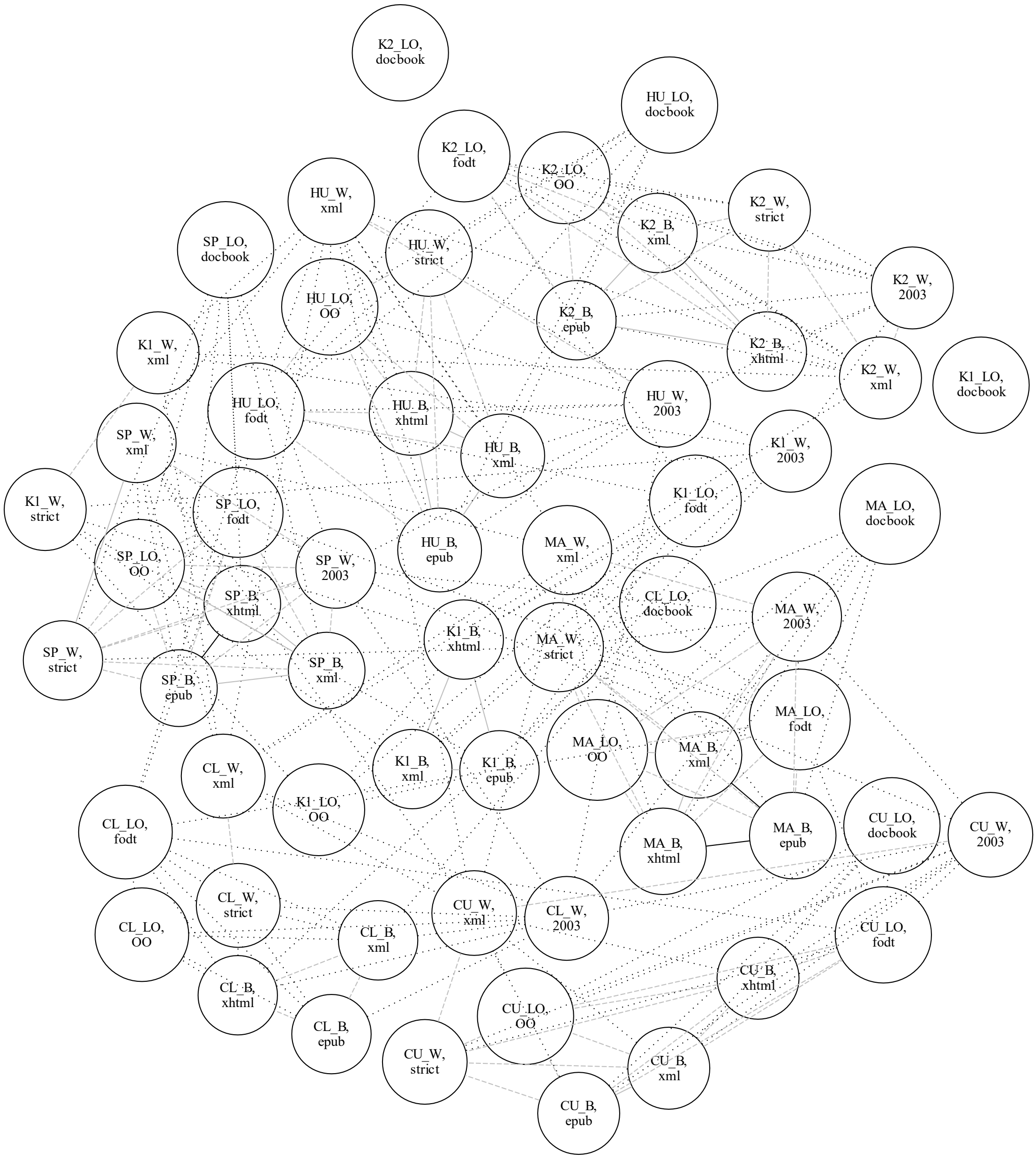
Notice, however, that this is the only of the 10 set- and bag-based measures in which there is a visible gap in the range, with pairs of thematically similar documents above the gap and pairs of thematically dissimilar documents below it.[41]

Identifying similar markup

In order to experiment with similarity measures on documents encoded in different markup vocabularies, we took three copies of each document in testbed 3 from the Balisage proceedings web site (the Balisage XML, the XHTML found in the epub version of the paper, and the HTML served to browsers) and translated them into six other formats using Libre Office (saving as Flat XML ODF document, docbook.xml, and Open Office XML) and Microsoft Word (saving as Word XML, Word 2003 XML document, and Strict Open XML document).

Applying the various methods for set- and bag-based measures directly on these variously marked up documents, however, did not provide useful results, as exemplified by this plot of the Jaccard similarity on bags:

Figure 7: 7 Balisage articles, each in 9 formats, Jaccard measure on bags



There are no useful clusters here, no groups of documents clearly closer to each other than to documents outside the group. Many of the adjacent pairs are indeed versions of the same paper, but not all, and the versions of a given paper do not form identifiable clusters.

It appears (but it is hard to think of good ways to test this hypothesis) that what has happened here is that the frequent occurrence of particular element type names (and namespace names?) has caused documents using the same XML vocabularies to be classed as similar, because of the high frequency of common items for things like paragraphs and titles. This does not completely swamp the effect of the text nodes, but it does obscure it.

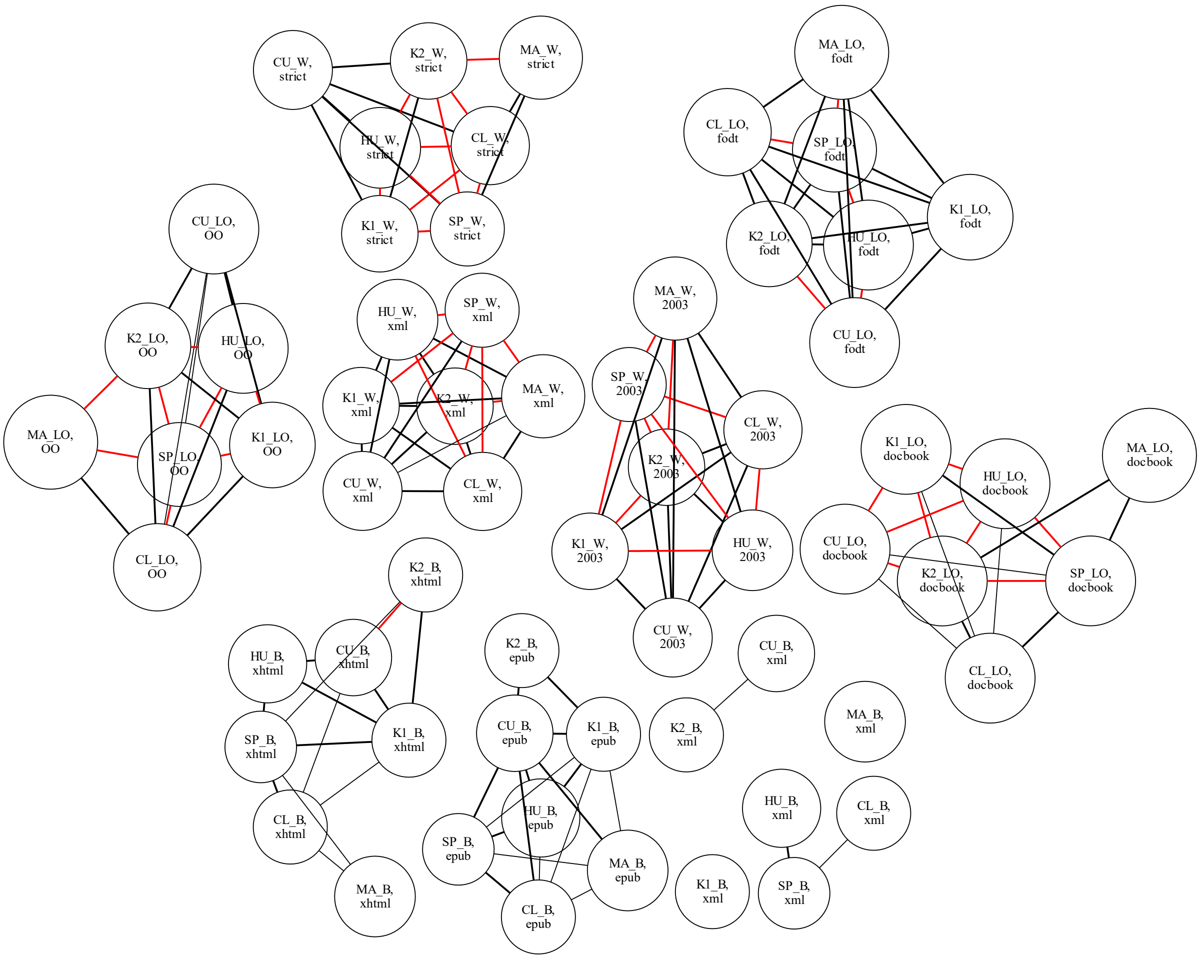
In order to try to separate the effect of the documents' character data content from the effect of the markup itself, we created various derived forms from each XML document, filtering the documents in various ways to emphasis different aspects of the markup structure in the documents. In all of them, comments, processing instructions, teiheader, script and style elements as well as attribute-value pairs were filtered out. In the discussion below, we will refer to the various derived forms using the following labels:[42]

- **m**: Both element nodes and text nodes are included in the bag or set used to model each document.
- **e**: Element nodes are included in the bag or set used to model each document; text nodes are leveled to the same dummy value. (So the model distinguishes between presence and absence of text nodes, but not between different text nodes.)
- **p**: The set and bag models are populated using parent-child pairs (e.g. body/div, div/h1, div/p, etc.) — a simple application of n-grams.

In retrospect, the results were perhaps predictable; we confess nevertheless to having been surprised by them.

For example, the Szymkiewicz-Simpson similarity measure on sets of the p transform produces the following plot:

Figure 8: 7 Balisage articles, each in 9 formats, transform p, Szymkiewicz-Simpson measure on sets



The clusters correspond almost exactly to the nine differently marked up versions of the seven documents. Similar plots were produced by several of the set- and bag-based similarity measures using transforms p and e.

It is reassuring, in its way, to find that when similarity measures are applied to sets or bags of element type names, the measures successfully classify documents by the vocabulary used to mark them up. It is, however, not a particularly compelling use case for quantitative measurements of document similarity to use them to do what can be done easily by just examining the root elements and namespace declarations of the various documents.

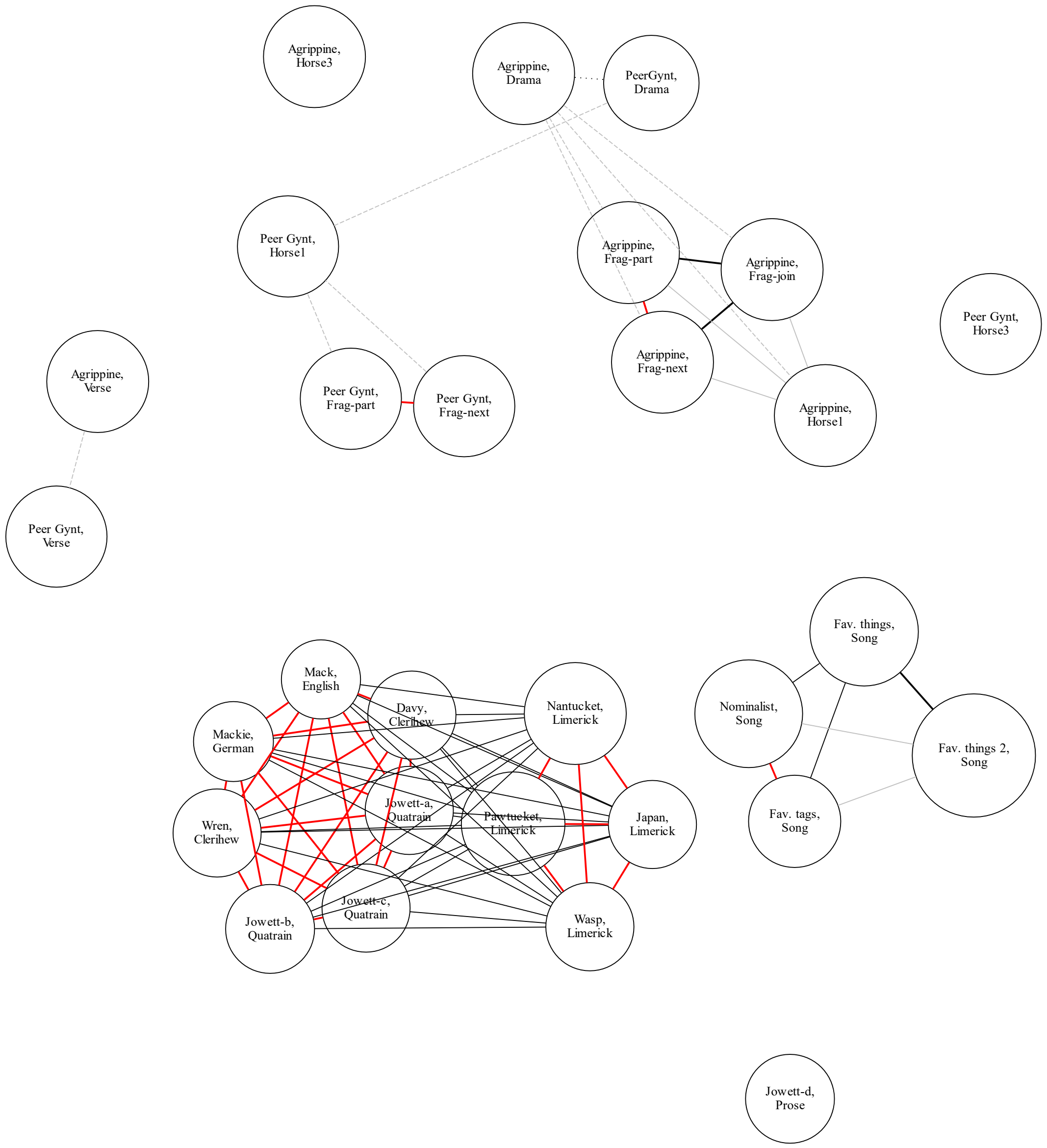
Identifying structurally similar documents

Our final experiment involved a collection of 29 short documents and document fragments tagged in TEI, consisting of

- some limericks (with five lines per line group, and one line group per poem)
- some clerihews and other poems with the same structure (four lines per line group, one line group per poem)
- some other verse (a song with words by Oscar Hammerstein and two parodies, with two line groups of four lines each and one of five lines)
- a short prose text
- some short fragments of verse drama from the MLCD Overlap Corpus, tagged in various ways: several versions of a couple of lines from Henrik Ibsen's *Peer Gynt*, and several versions of some lines from Cyrano de Bergerac's *Agrippine*.

Since all of the documents are using the same parent vocabulary, we hoped it might be possible to use markup-sensitive similarity measures to detect useful groupings of these documents. Jaccard similarity on bags on the e transform produces the following plot:

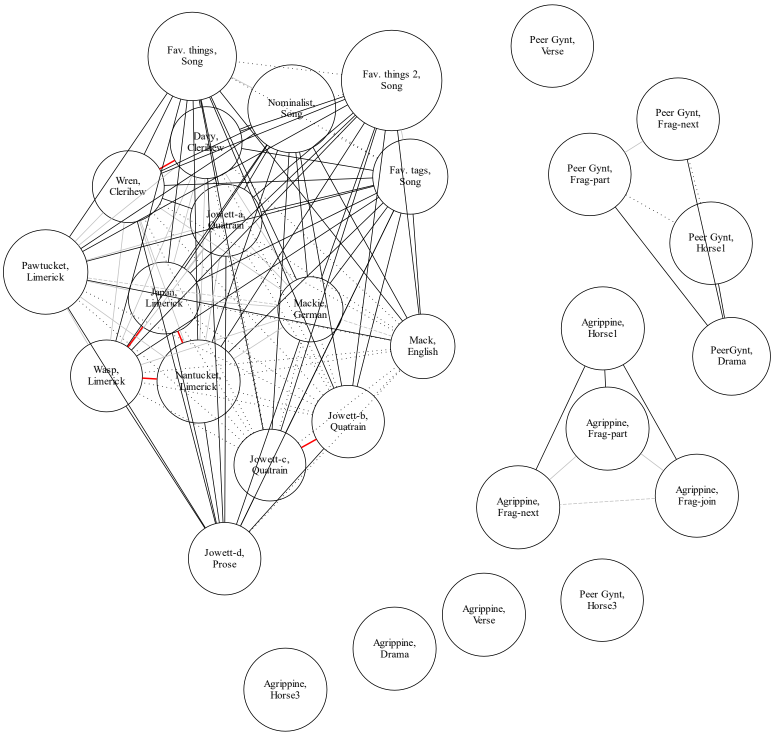
Figure 9: 29 TEI documents, transform e, Jaccard measure on bags



As can be seen, there are links between the verse and drama versions of Peer Gynt and Agrippine, the four songs form a separate cluster, the limericks seem to form a sub-cluster within the other poetry cluster, and the only prose text in the collection does not come out as similar to any of the other documents. (On the other hand, it is surprising that the same goes for the Horse3 versions of Peer Gynt and Agrippine.)

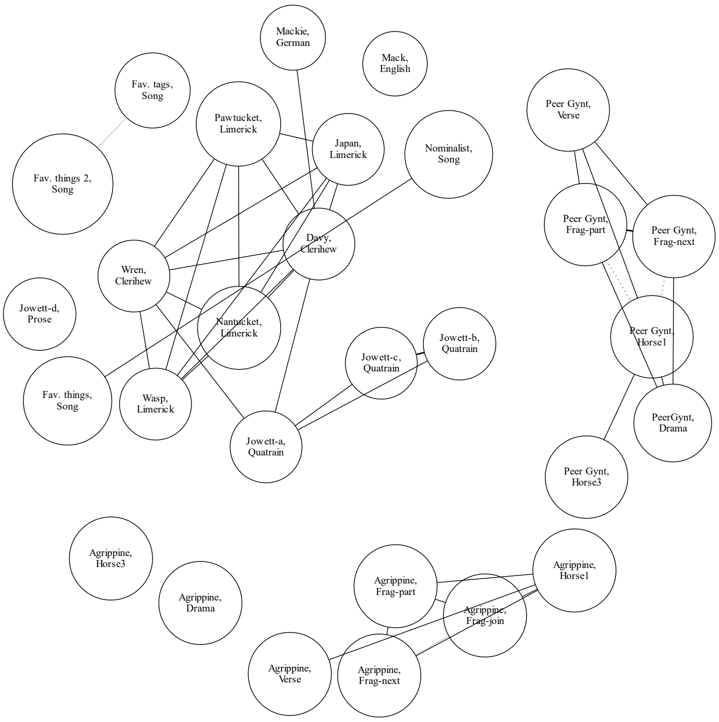
A Jaccard similarity on bags, with text nodes leveled and with n-grams of length 1 to 5 shows slightly different clustering: here, the versions of Peer Gynt cluster with each other, as do the versions of Agrippine (with a few strays), but again all the non-dramatic verse is in a single cluster.

Figure 10: 29 TEI documents, transform e, Jaccard measure on bags of 1-5-grams



A similar plot for the same model with text nodes not leveled clusters versions of the same or similar texts more tightly (as might be expected), without completely dissolving the markup-based clusters:

Figure 11: 29 TEI documents, transform m, Jaccard measure on bags of 1-5-grams



Conclusion

Since the introduction of writing, transcription has been the method by which documents have been produced which are “as similar as possible”, or even “so similar as to be, for practical purposes, identical” to other documents. From a philosophical point of view, of course, no document can be identical to any other document, and whether a given kind or degree of similarity suffices for particular purposes will invariably depend upon the purposes. For purposes that depend upon the linguistic content and meaning of a document, transcription appears to provide an effective culturally developed standard of maximal similarity.

Our earlier work on transcription has led us to postulate that in principle an error-free transcription possesses the property of t-similarity with its exemplar. T-similarity can be decomposed into the properties of purity, completeness, and type identity with the exemplar. That is, a t-similar transcription neither adds nor omits nor changes any of the types present in the exemplar, whether atomic or composite. The basic rules of t-similarity may be seen to provide a rationale, or a common theoretical underpinning, of the various similarity measures we have discussed.

Standard similarity measures for sequences exhibit concern for precisely the addition, omission, or replacement of members of the sequence, and can be interpreted as measuring similarity between two sequences by counting the number of points at which the two sequences fail to be perfectly t-similar. Many different similarity measures can be constructed by ignoring composite tokens (e.g. the sequence of words in a document, or the arrangement of words into paragraphs and paragraphs into sections, etc.) either selectively (as sequence measures ignore higher-level types) or comprehensively (as bag and set measures ignore all composite types and work only with word types or word occurrences).

Further variants (which we have not discussed here) can be constructed by weighting different additions, omissions, and changes differently; for subject-matter retrieval, for example, the replacement of one synonym by another should ideally not lead to a lower similarity measure. Yet more variants can be created by increasing ingenuity in the calculation of different weights for different substitutions, based for example on distributional semantics.

We can if we wish make similarity measures sensitive to markup structures by defining graph- or tree-edit similarity measures analogous to the sequence edit distances like Levenshtein distance and Damerau/Levenshtein distance. There is useful research being done on this topic. Unfortunately for those with short-term needs for practicable solutions, the calculation of edit distances on trees and graphs appears to be expensive and the distances measured do not appear always to correlate with our intuitions about the relative similarity of various pairs of marked up documents.[43]

For some applications, including the detection of genetic or thematic relations among documents, markup sensitivity appears not to be needed. When it is needed, some sensitivity to markup can be achieved by using bag and set similarity measures on documents modeled as bags or sets of n-grams, where various markup constructs (element nodes, attribute nodes, etc.) and various adjacency relations (parent/child, adjacent sibling, ancestor/descendant, etc.) are used to create the n-grams. If the markup is of particular interest, it can be helpful to filter out the words of the text, so that the markup items are the only members of the set or bag.

Applying document similarity measures to documents encoded in different vocabularies is unlikely to reveal much beyond the fact that the documents use different vocabularies. But when documents are encoded in the same vocabulary, n-grams constructed to include markup can be used to recognize structural similarities between documents.

As we have seen, simple measures may be surprisingly effective. But until we have persuasive ways to quantify similarity between trees and graphs that go beyond standard similarity measures operating on tokens and types of their serializations, document similarity will remain an open problem.

References

[Damerau 1964] Damerau, Fred J. 1964. “A technique for computer detection and correction of spelling errors”. *Communications of the ACM* 7.3 (March 1964): 171-176. doi:https://doi.org/10.1145/363958.363994.

[Douven and Decock 2010] Douven, Igor, and Lieven Decock. “Identity and Similarity”. *Philosophical Studies* 151.1 (2010): 59-78. dio:https://doi.org/10.1007/s11098-009-9415-5.

[FRBR] International Federation of Library Associations. *Functional Requirements for Bibliographic Records*. IFLA Series on Bibliographic Control 19. Munich: K. G. Saur, 1998.

[Goodman] Goodman, Nelson. “Seven strictures on similarity”. *Problems and Projects*. Indianapolis: Bobbs-Merrill, 1972.

[Huitfeldt, Marcoux and Sperberg-McQueen] Huitfeldt, Claus, Yves Marcoux and C. M. Sperberg-McQueen. “Extension of the type/token distinction to document structure”. Presented at Balisage: The Markup Conference 2010, Montréal, Canada, August 3 - 6, 2010. In *Proceedings of Balisage: The Markup Conference 2010*. Balisage Series on Markup Technologies, vol. 5 (2010). doi:https://doi.org/10.4242/BalisageVol5.Huitfeldt01.

[Huitfeldt] Huitfeldt, Claus. 2006. “Philosophy Case Study”. In *Electronic Textual Editing*, ed. Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth. New York: MLA 2006, pp. 181-96.

[Jaccard] Jaccard, Paul. “Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines”. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 (1901): 241-272. doi:https://doi.org/10.5169/seals-266440. [Not seen.]

[Levenshtein] Levenshtein, Vladimir I. “Binary codes capable of correcting deletions, insertions, and reversals”. *Soviet Physics Doklady* 10.8 (1966): 707–710.

[Peirce] Peirce, Charles Santiago Sanders. “Prolegomena to an apology for pragmaticism”. *The Monist* 16 (1906): 492-546. doi:https://doi.org/10.5840/monist190616436. Reprinted vol. 4 of C. S. Peirce, *Collected papers*, ed. Charles Hartshorne and Paul Weiss (Cambridge, MA: Harvard University Press, 1931-58).

[Van der Meulen and Tanselle] Van der Meulen, D. L. and Tanselle, G. T. “A system of manuscript transcription”. *Studies in Bibliography*, 52: 201–12 (1999).

[Wittgenstein] *Wittgenstein's Nachlass: The Bergen Electronic Edition*. Oxford University Press, 2000, ISBN-10:0192686917.

[1] Most or all legal systems have rules for authentication of copies of passports and other legally important documents. Typically, these rules require a person in some professional role to confirm having seen both the copy and the original, and to testify (in writing) that the one is a true copy of the other.

See https://www.gov.uk/certifying-a-document : “Take the photocopied document and the original and ask the person to certify the copy by: writing Certified to be a true copy of the original seen by me on the document; signing and dating it; printing their name under the signature; adding their occupation, address and telephone number”.

In less rigidly controlled situations, however, we may be content to conclude from the identity of the name, the time stamp and the size of two computer files that they are copies of the same document. The identity of these features are of course neither necessary nor sufficient conditions for document identity. File comparison tools may automatically verify that two document files both exemplify the same sequences of bits. This is usually a sufficient, though far from always a necessary condition for the effective identity of the two documents.

[2] We use the word “object” in a very wide sense. Objects may be physical, concrete, continuous, dispersed, singular, plural, universal, abstract, and so on. In general, anything to which one may attribute a property is an object.

[3] It should perhaps be noted that Leibniz postulated that difference in physical location alone cannot suffice to establish a claim that two things are in fact distinct things; this is part of his doctrine of the identity of indiscernibles (see next note). We are unable either to follow him in this postulate or to provide a principled reason to reject it.

[4] Here the expressions $x=y \Rightarrow (Fx \Leftrightarrow Fy)$ and $(Fx \Leftrightarrow Fy) \Rightarrow x=y$ are to be understood as holding for all objects x and y and all predicates F. The identity of indiscernibles, also known as “Leibniz's law”, has been (and still is) disputed. The indiscernibility of identicals, by contrast, is generally accepted, perhaps because it appears to follow from the law of excluded middle.

[5] It may be objected that, thus construed, identity is not a relation between objects at all.

[6] Some famous examples: The ship of Theseus, paradoxes of personal identity and of constitution.

[7] A recent attempt to provide a reference model for some of the many different meanings commonly attached to words like *book* is given by FRBR.

[8] When we relate to digital documents, there is a sense in which we relate to physical objects, e.g. particular portions of data carriers with certain patterns of electromagnetic properties. But we rarely think of them in those terms, probably because they are not directly perceivable. What is directly perceived, like a book, is what we see on the screen (or hear from the loudspeakers). Alternatively, we may see it this way: a computer automatically, instantly and reliably creates new physical copies of the document, so we do not need to think about the relation between the physical representations on disk, in cash, or in screen memory, or the pixels on the screen, the projector etc.

That may be why people sometimes say about digital objects (document files), not that they are documents, but that they represent (or even instantiate) documents, which seems to suggest that the digital document itself is an abstract object with physical instantiations.

Note, for comparison, that we do not (normally) say that a book instantiates or is instantiated by a document, -- the book is the document, it is a physical object and thus cannot have instantiations.

The transition from traditional, “physical” to “digital” documents challenges our ideas about documents and text in ways we may still not quite have been able to make out. We tend to think of physical documents like books as directly perceivable and stable objects, in contrast to digital documents which cannot be directly perceived unless they are displayed on paper or screen, and which seem to be transient as they vanish when we turn off the device or look different when we adjust some parameters. But our access to the textual content of a book, too, requires us to pull it off the shelf and open it, it may have changed or even have been replaced since we last opened it, we depend on light for its readability, and so on... Considered as a perceivable and relatively stable portion of physical reality it fares no better than the digital document, at least in principle, though admittedly we rely on less robust and universally available technology in the one case than in the other.

[9] It should be noted at this point that nodes as defined by the XPath data model XDM are always tokens in nature; XDM defines no way to group tokens as being of the same type in Peirce's sense (XDM nodes do have types, but none of the various types attributed to them is a type in Peirce's sense). It follows that any similarity measure intended to apply to nodes in an XML document must define its own ways of reducing node tokens to types; it is far from obvious at this point which ways will work best for which purposes.

[10] by transcription we mean the effort to report—insofar as typography allows—precisely what the textual inscription of a manuscript consists of. [Van der Meulen and Tanselle, p. 201]

[11] A replica of a clay tablet may itself be a clay tablet, and thus very similar to the original tablet. Even the most accurate and faithful transcription of a clay tablet, on the other hand, will in most respects not be similar to the original clay tablet at all, and certainly not look like one. Transcription is about another kind of similarity.

[12] As already explained, although the type-token distinction is usually applied primarily to lower-level linguistic units like letters or words, the distinction also applies at higher levels, such as sentences, paragraphs, or whole documents, which are then seen as composite tokens of composite types.

[13] T-similarity as such only says something about the similarity relation between documents. It does not capture other salient facts of the relation between transcript and exemplar, for example that E must precede T in time, that E has served as evidence for T, that T has been made on the basis of E and with certain intentions and aims, etc. It is also immaterial to t-similarity whether the documents in question are hand-written, printed, digital, or carved on clay tablets.

Like any formally defined similarity, t-similarity is context-dependent: the claim that two documents are t-similar entails assumptions about the type system(s) used and various details of the transcription's practice. In consequence, the reflexivity, symmetry, and transitivity of t-similarity hold only under certain circumstances, the details of which would take us too far afield in this paper. The main point, however, still holds: t-similarity is not the same relation as transcription.

[14] Alban Berg's opera is called *Wozzeck* because that was how the name was read by the editor who first published the material. Later editors, guided in part by documentary records concerning the historical figure who inspired Büchner's play, have read the name *Woyzeck*.

[15] See the Wikipedia article on metrics.

[16] Diff is strictly speaking not a similarity measure, but a tool for identifying modifications necessary in order to transform one document to another. A count of the number of modifications identified may however be used as a similarity measure. In practice, diff programs make some compromises in the interests of speed of execution, so the sets of changes they show are not guaranteed to be absolutely minimal.

The *Hamming distance*, defined in 1950 by the Bell Labs researcher Richard Hamming, was designed to find ways of ensuring that messages encoded in a particular way would be sufficiently dissimilar to ensure that they would not be mistaken for each other even if the encoding were damaged in transmission. For any strings of characters A and B, with the same length, the Hamming distance between them is the number of positions in the string at which A and B have different characters. The Hamming distance does often not match our intuitive sense of how similar two sequences are, and it is undefined for strings of unequal length. (In practice, if people want to compute Hamming distances for strings of different lengths, they pad out the shorter string with some character not present elsewhere.) Therefore, it seems ill suited for measurement of document similarity.

[17] As explained above, by “similarity” we mean “normalized similarity” when we say, for short, “Levenshtein similarity” and “Hamming similarity”. We use pairs of strings of equal length as examples here, because the Hamming distance is not defined for strings of unequal lengths.

[18] This account also suggests why the Hamming distance seems less satisfactory as an account of similarity: it does not recognize the categories of impurity and incompleteness [insertions and deletions], only that of type difference [replacement or substitution].

[19] As can be seen from the *Merkel* / *Markle plagiarism* / *plagiarsim* examples, the Levenshtein distance treats transpositions like *le* / *el* as adjacent substitutions; when the two letters are the same, however, the Damerau/Levenshtein distance counts only one change, not two. (So the Damerau/Levenshtein similarity for *Merkel* / *Markle* is 0.67, compared to a Levenshtein similarity of 0.5; for *plagiarism* / *plagiarsim* it is 0.9, compared to 0.8.)

[20] Also known as the *Jaccard index* or the *Jaccard similarity coefficient* (*coefficient de communauté*), defined by the scholar Paul Jaccard in 1901 [Jaccard].

[21] For example, if we wish to measure the similarity of the words *Merkel* and *Markle*, we have A = {e, k, l, M, r}, B = {a, e, k, l, M, r}, A∩B = {e, k, l, M, r}, A∪B = {a, e, k, l, M, r}, and thus a Jaccard similarity of 5/6, or 0.83. The details for the other entries in the table below are left as an exercise for the reader.

[22] A bag of objects is like a set in that it is unordered, but unlike a set in that one and the same object can occur more than once. Bags are sometimes called multisets. A bag is also like a sequence, in that a given type may occur more than once; but like items shaken together in a bag, the members of a bag of characters do not have any meaningful order.

[23] The method of counting the members of the intersection and union must be adjusted to account for the shift from sets to bags: the intersection contains the lower count for each type present in both inputs, the union contains the higher count for any type present in either or both.

[24] In this case, we have padded the shorter string to the same length as the longer before calculating a normalized Hamming distance.

[25] We have not been able to find this measure in the literature we have consulted so far, but we guess the likelihood is low that it hasn't been thought of before. The authors would be grateful for references.

[26] See https://en.wikipedia.org/wiki/Overlap_coefficient. Sometimes called the overlap coefficient or index. We sympathize with readers who have reported that they find the term “overlap” misleading: the Szymkiewicz-Simpson similarity is 1 whenever one of the two collections is a sub-collection of the other.

[27] See https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient.

[28] This variant is also known as the Otsuka-Ochiai coefficient. See https://en.wikipedia.org/wiki/Cosine_similarity.

[29] In some language corpora it would be represented as a sequence of 38 tokens (34 words and 4 punctuation marks). (These figures assume that we are leaving the title and author attribution out of the comparison, which often goes without saying, but should probably not.)

[30] Treating each text node or verse line as a token may seem far-fetched at first glance, but it would quite possibly be the best basis for comparing the similarity of two sonnets in Raymond Queneau's book *One hundred billion poems*. Similarity measures often vary with context!

[31] Some word tokens may be split over markup boundaries, but for purposes of constructing the set or bag in question here, that need not trouble us: we can include both the uninterrupted word and the interrupting markup as elements of the bag.)

[32] We cannot, for this purpose, usefully use nodes as defined by XDM: since no XDM node can occur in more than one document, it follows that the intersection of the node sets of any two documents will be empty, which means that most set-based similarity measures will return a value of zero. But we can abstract away from node identity a little bit by defining an equivalence relation on nodes (whose equivalence criteria can be chosen to taste) and working with the resulting equivalence classes.

[33] Actually, there are many pairs of relations that can be used: one could take document order and preceding-sibling, or previous-node and descendant. The key observation is that for an ordered tree, a set of nodes and two relations on those nodes suffice. That suggests performing similarity measurements on XML documents by applying set- or bag-based measures to the set of nodes, and the set of pairs in the three defining relations.

[34] The articles are:

- DGM: Deutsche Gesellschaft fuer Muskelkranke; 2019-09-21
- DAK: Donald E Knuth; 2019-10-23
- GF_a: Gottlob Frege; 2009-08-05,
- GF_b: Gottlob Frege; 2016-12-30>
- GF_c: Gottlob Frege; 2020-02-06
- H: Halva; 2020-03-02
- LW_a: Ludwig Wittgenstein; 2010-12-21
- LW_b: Ludwig Wittgenstein; 2014-07-15
- LW_c: Ludwig Wittgenstein; 2020-03-01
- NL_a: Niklas Luhmann; 2018-10-01
- NL_b: Niklas Luhmann; 20120-03-28

- NW: Niklaus Wirth; 2020-03-04

[35] The siglia of these paragraphs are: Ms-102,1r[2], Ms-102,2r[3], Ms-104,118[6], Ms-104,12[13]et13[1], Ms-104,3[13], Ms-104,47[6], Ms-105,32[3]et34[1], Ms-113,117v[3], Ms_103,83r[3], Ts-202,19r[5], Ts-202,21r[1], Ts-202,28av[1], Ts-202,28av[1]-relocated, Ts-208,54r[6], Ts-209,74[2], and Ts_204,30r[1].

[36] The seven articles are:

- CL: Ashley M. Clark: “With One Voice: A Modular Approach to Streamlining Character Data for Tokenization” https://www.balisage.net/Proceedings/vol23/html/Clark01/BalisageVol23-Clark01.html
- CU: Autumn Cuellar and Jason Aiken: “The Ugly Duckling No More. Using Page Layout Software to Format DITA Outputs” https://www.balisage.net/Proceedings/vol17/html/Cuellar01/BalisageVol17-Cuellar01.html
- HU: Claus Huitfeldt, Yves Marcoux, C. M. Sperberg-McQueen: “Extension of the type/token distinction to document structure” https://www.balisage.net/Proceedings/vol5/html/Huitfeldt01/BalisageVol5-Huitfeldt01.html
- K1: Eliot Kimber: “DITA Grammar Customization. Enabling controlled grammar extension for loosely-coupled interchange and interoperation” https://www.balisage.net/Proceedings/vol24/html/Kimber02/BalisageVol24-Kimber02.html
- K2: Eliot Kimber: “Hyperdocument Authoring Link Management Using Git and XQuery in Service of an Abstract Hyperdocument Management Model Applied to DITA Hyperdocuments” https://www.balisage.net/Proceedings/vol15/html/Kimber01/BalisageVol15-Kimber01.html
- MA: James David Mason: “Do we really want to see markup?” https://www.balisage.net/Proceedings/vol23/html/Mason01/BalisageVol23-Mason01.html
- SP: C. M. Sperberg-McQueen: “Thinking, wishing, saying” https://www.balisage.net/Proceedings/vol23/html/Sperberg-McQueen02/BalisageVol23-Sperberg-McQueen02.html

CU, K1, and K2 are classified by the Balisage website to be related to the same topic, “DITA”.

[37] The sigla of the three paragraphs are: 1) Ms_103,83r[3], 2) Ms_104,118[6], and 3) Ms_104,3[13].

[38] The red links connect documents with similarity 1.0, black bold links connect documents with similarity greater than 0.9, black unbolded links 0.8, solid gray 0.7, dashed grey 0.5, and dotted grey 0.3. Distances between documents not connected by links are used for layout purposes, but the links are not drawn.

The layout is performed by the Graphviz *neato* layout engine, which accepts suggested lengths for edges connecting nodes in the graph, and adjusts them as needed using a physical model involving springs, which performs a kind of multidimensional scaling. The distances cannot be guaranteed accurate, since there is no guarantee that the distance measure used describes a two-dimensional Euclidean space. But as may be seen Graphviz does a creditable job.

[39] The careful observer will note that more connections are shown in the second diagram than in the first: this reflects partly the fact that the symmetric subsumption measure will always be higher than the Jaccard measure, and partly that the one measure is a set measure and the other a bag measure.

[40] And how can we be certain of these results? In the case of the Wittgenstein Nachlass, we have considerable independent evidence on the genetic relationships between a very large number of paragraphs.

[41] In this case, the range for all 21 pairs is 0,1114-0,2925, and the values of the three first pairs are 0,2925, 0,2625, and 0,2198. The corresponding values for the Szymkiewicz-Simpson similarity on sets are: range 0,2081-0,4226, three first pairs 0,4226, 0,3123, and 0,2988.

[42] We illustrate here by showing transforms m, e, and p for the following document:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader> ... </teiHeader>
 <text>
  <body>
   <lg n="a">
    <l n="1">Sir Humphy Davy</l>
    <l n="2">Abominated gravy.</l>
    <l n="3">He lived in the odium</l>
    <l n="4">Of having discovered sodium.</l>
   </lg>
<!-- Comment -->
<?Processing instruction ?>
  </body>
 </text>
</TEI>
```

Transform m:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <text>
      <body>
         <lg>
            <l> Sir Humphy Davy </l>
            <l> Abominated gravy. </l>
            <l> He lived in the odium </l>
            <l> Of having discovered sodium. </l>
         </lg>
      </body>
   </text>
</TEI>
```

Transform e:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <text>
      <body>
         <lg>
            <l> E </l>
            <l> E </l>
            <l> E </l>
            <l> E </l>
         </lg>
      </body>
   </text>
</TEI>
```

Transform p:

```
TEI--text
text--body
body--lg
lg--l--E
lg--l--E
lg--l--E
lg--l--E
```

[43] This is not terribly surprising, perhaps, given that the same problem arises for sequence measures, as illustrated by the pair *abcdefghij*, *fghijabcde*, created by selecting a single contiguous substring and moving it, but assigned a normalized Levenshtein distance of 1, and a normalized Levenshtein similarity of 0.

Author's keywords for this paper:

Document similarity; Transcription; Levenshtein; Jaccard

Claus Huitfeldt

Associate Professor

University of Bergen

<**Claus.Huitfeldt@uib.no**>

Claus Huitfeldt is Associate Professor at the Department of Philosophy of the University of Bergen, Norway. He was founding Director (1990-2000) of the Wittgenstein Archives at the University of Bergen, for which he developed the text encoding system MECS as well as the editorial methods for the publication of Wittgenstein's Nachlass - The Bergen Electronic Edition (Oxford University Press, 2000).

C. M. Sperberg-McQueen

Founder and principal

Black Mesa Technologies LLC

C. M. Sperberg-McQueen is the founder and principal of Black Mesa Technologies, a consultancy specializing in helping memory institutions improve the long term preservation of and access to the information for which they are responsible.

He served as editor in chief of the TEI Guidelines from 1988 to 2000, and has also served as co-editor of the World Wide Web Consortium's XML 1.0 and XML Schema 1.1 specifications.