Support for Discursive Engagement
Reference and Identity on the Web

Brian Cantwell Smith

Version 0.3 — 2011 April 5

Early sketch—by no means even a full draft!—of a paper, written for discussion at iTable, Faculty of Information, University of Toronto. Not for citation or reference. The project is being developed in collaboration with other members of Amanda/Group of N, to whom many thanks. Infelicities of presentation are my own. Comments welcome: please send to brian.cantwell.smith@utoronto.ca.

1 · Introduction

Digital technologies have had a profound influence on intellectual inquiry, human interaction, and documentary practices—what in general I call **discursive engagement.** Think of the sheer diversity of digital forms with which we have become familiar: *email, distribution lists, blogs, electronic publication*, ubiquitous *.doc* and *.pdf files, multi-threaded forums* and *bulletin boards, web resources* (such as Google docs), *"track changes"* editing protocols in Microsoft Word, sundry *social media, data bases*, the *"cloud,"* etc.—to say nothing of the background ability to compose, edit, adjust, layout, annotate, and disseminate documents in digital form. By stitching together a motley bricolage of these new forms, based on a plethora of protocols, standards, and formats, we have reconfigured, in unprecedentedly rich and interconnected form, the fabric of human creative expression.

Moreover, the relentless pace of technological innovation and socio-technical reconfiguration continues unabated. No matter how consequential, accomplishments to date are meagre compared to what is possible. In fact it seems undeniable that history will affirm the triumphalists' claim: we have barely begun to unleash the web's potential.[1]

Less widely agreed is the status of the current state of the art. In this paper, I will argue that some of the deepest assumptions on which the architecture of the web is currently based are blocking our ability to make radical forward progress. Proper support for discursive engagement, I will argue, will not emerge by making incremental adjustment to existing formats, protocols, and standards. Instead, it will require rethinking, at the most fundamental level, some of the most basic concepts and categories we use to understand, design, and use the internet.

1a — Assumptions

I particularly want to question perhaps the most pervasive existing *mythos* about

[1] Or to put it a bit more theoretically soundly: society has barely embarked on the transformation of persons and societies that will ultimately be catalysed and enabled by the transformation of the material substrate of creative expression from the registration of marks on paper to configurations of digital arrangement.

the nature of the web: that it comprises a *connected fabric of (hyper)linked documents*.[2] Three issues underlying this mythos are particularly important to the structure of discursive engagement. Two are positive assumptions on which its architecture has been built:

1. **Objects and Identity**: The fundamental entities out of which the existing web is built have been forged on what I will label a *classical notion of an object*—i.e., on an object as an intrinsically singular[3] (though not necessarily atomic) entity, typically of some type or category or class, exemplifying properties, standing in relations, collected into sets or groups. The most visible form of internet object may be the *document*, but there are others: *files*, *web sites*, *resources*, *code*, etc.[4] Crucially, and increasingly, the net is also *active*—not just in hosting dynamic media, but serving as a platform for action and interaction. On an "objectivist" ontology, those dynamic phenomena can themselves be taken to be a form of temporally-changing object.

   The most important mark or characteristic of objects—the property that most tellingly betrays how objects are treated, what assumptions they make, what consequences follow from their characterisation, etc.—has to do with their **identity**: what it is that distinguishes one object from two, what warrants a claim that a given object is *the same* as or *different from* another. Issues of identity will take centre stage in what follows. Roughly, I will argue: that the forms of identity needed to understand the world of documents in particular, and human discursive practices more generally, radically transcend anything that has been, or can be, imagined from within a classical "object-oriented" ontological model.

2. **Connective tissue:** The most visible conception of the connective tissue tying the web together is that of a *link*—the ubiquitous distinguished (textual) form, architecturally supported, which may be used within a file, document or other web or net object in order to target or point to, and thereby provide access to, another network entity. Of links with which we are accustomed, IP addresses, URLs, and email addresses are among the most common, though we are increasingly seeing other forms, such as unique identifiers (URIs) and proper names. Links provide a crucial addition to the more direct form of connectedness that arises from *inclusion*, as for example with quoted fragments, enclosed emails, files within online directories, etc., where one "piece" of networked structure, or (crucially) a copy, is included within a composite other.

[2] Or even: as interlinked *data*, as Tim Berners-Lee has recently argued («ref TED talk»).»

[3] By 'intrinsically singular' I mean that the object's identity as a (singular) entity is taken to be an inherent or intrinsic property of it. Identity, that is to say, is assumed to "inhere" in the object itself. This is the most fundamental criterion of object-hood that is rejected in the proposal presented in §3.

[4] These latter can all be classified as types of document—but I will not belabour that here.

> Because of their prominence and architectural support, it might be thought that inclusion plus links (especially IP addresses and URLs) underwrite all of the internet's interconnectedness—that, together, they constitute the "warp and woof" of the web.[5] I do not believe this is true. Below I will argue that links, especially, capture only a small fraction (as little as 10%?) of the web's interconnectedness. Much (40%?) remains buried in uninterpreted natural language text; another large fraction (another 40%?) is codified in diverse, ad-hoc, non-generalizable protocols and mechanisms that remain mostly inaccessible to search and discovery, except within the very particular (and typically local) contexts and situations for which they are designed and in which they are used (such as message quotations within email and forum posts, changes tracked in Microsoft Word, etc.).
>
> In place of links, I will recommend reconceiving of the web's interstitial connectivity in terms of a constructive version of *reference*—specifically, a variety I will call **registration**.

By 'discursive engagement' I include myriad forms of human expression and interchange, including the indissoluble mix, at arbitrary scales, of creative originary work and commentary and response to the works of others—from short email messages and forum posts through blogs, reviews, citations, exegesis, documentation, drafts, papers, manuscripts, and books, up to and including the full interconnected texture of scholarship. A third topic, as important to this range of practices as that of object identity and reference, is not so much one on which the classic conception of the web takes a *problematic stand*, but rather one to which it *fails to do justice*. In particular, our classic conception of links, documents, content, identity, etc., provides us with neither the wherewithal nor the "space" for adequate treatment of:

3. **Multiple registration:** The fact that all objects, phenomena, entities, situations, etc., can be understood as *consisting of*, or *being intelligible at*—or, as I will say, can *be registered*—in multiple, cross-cutting ways, in terms of different dimensions or aspects, as being constituted of an assemblage of parts organized according to multiple, cross-cutting mereological "parses," etc.

Thus a student paper—to take a simple example, and for the moment setting issues of identity aside—can be registered as consisting of: (i) a sequence of *characters*; (ii) a sequence of *sections*, perhaps preceded by a title and author, and extended with a list of references—in which each section consists of a number of *paragraphs*, each in turn comprising a run of *sentences*, those composed of *words*, etc.; (iii) a sequence of *pages* (at least when printed or presented or rendered on pa-

[5] I am assuming—contrary to fact!—that the terms 'internet' and 'web' are synonymous, which they are not. For purposes of this paper, however, the differences are not germane.

per[6]), each comprising a list of lines, each (again) comprised of a sequence of characters (if it is rendered in a character-based language); and so on.

The fact that documents support multiple registration is not contentious. What is problematic are the implications of this fact for architectural support. In the general case, I will argue—i.e., for an arbitrary unit or segment of discursive exchange—no one registration will necessarily be able to be dubbed as the *original* or *primary*, with the others being *derivative* or *secondary*. Nor, relatedly, can we in general assume that any particular registration is fundamental (even though it may seem as if the "sequence of characters" registration has foundational status in current document systems[7]), or that it any necessarily be *derived* from the other. Neither are relationships between and among registrations necessarily forms of "coarse-graining," ultimately grounded on a fixed fully-specified "bottom level" registration. Nor, as is evident from the example of pages/lines/characters, does "fixity" in one registration necessarily imply "fixity" in another, since "one and the same document," at least according to a common view of document identity, can be "rendered" or "paginated" in multiple diverse ways. And most significantly, to reach forward a bit, the ways in which a given entity is registered interacts with the appropriate forms of identity in terms of which that entity is individuated.

One way to understand this paper, therefore, is as an exploration of what will be required in order for us, as theorists and web architects alike, to deal appropriately and creatively with these three notions of *identity*, *reference*, and *(multiple) registration*, so as to provide maximally powerful and useful support for human discursive practice. The argument ultimately ends up arguing for a whole new metaphysical (ontological and epistemological) approach, based on a diagnosis of what are argued to be insurmountable problems our existing classical models. These are characterised as being of two overarching kinds: ontological and semantic. A word on each.

1b — Ontology

The deepest problem with our current understanding of the web, in my view, and of the architectures we have implemented and based it on, drive from what I described above as *classical ontology*: the picture of a world consisting of intrinsically singular discrete reidentifiable objects exemplifying properties and standing in relations. Interestingly, though it may seem the epitome of naïve realism, classical ontology, in the sense I am using the term, is strictly neutral as between realistic, idealist, constructivist, relativist, and other metaphysical stances (one can for example develop an idealist conception of intrinsically singular objects). More relevant for our purposes is its ubiquitous use in the design, analysis, understanding and use of formal systems. For example, it is universally assumed by formal logic. More pertinently, it underwrites all extant computational standards—from machine language and C++ up through TCP/IP to include URLs, URIs, XML, RDF, and

[6] The situation on digital screens is of course more complex; see below.
[7] This may seem to be true—but is not in fact the case.

a variety of proposals for structuring the semantic web.[8,9]

Classical ontology has served science well—e.g., for classical mechanics, for the mathematization of science, for the development of formal logic, etc. But in spite of sitting in the driver's seat for several centuries, classical ontology over-simplifies the nature of the world. And like all simplifications, there comes a time when it must be replaced. I will argue here that *one cannot do justice to the world of documents and human discursive expression* from a classical ontological viewpoint. What is required, rather, is an approach more consistent with the sorts of constructive epistemological and ontological commitment explored and embraced in cultural theory, feminist epistemology, science and technology studies, post-structuralism, and other branches of contemporary epistemology. As will be explored in §3, we need computational systems and calculi developed from the ground up in terms of such an overall epistemological/ontological viewpoint.

1c — Semantics

The second main group of issues relevant to the reconstruction to be proposed are semantical. A preparatory note on vocabulary.

Interestingly, computer science imported a great deal of its theoretical vocabulary from logic, including a large number of semantical terms: *symbol*, *identifier*, *reference*, *semantics*, *language*, *name*, *data*, *information*, etc. A classicist might therefore expect computer science to be a semantically rich discipline. Strikingly, however, for historically intelligible if nevertheless unfortunate reasons, the computational community has largely reconfigured and reconceptualized these classically semantical terms for mechanistic and/or computational "internal" purposes. As a result, it has become extraordinarily difficult for classicists and philosophers and others in the academy to talk, in ways that are intelligible within computational circles, about what they might think of as *genuinely semantical* issues. It seems that what one says is recognized and understood, because the words are familiar—but what one is actually saying gets lost in translation. The originary semantical issues themselves, of course (speaking as a classicist!) have not gone away; and some new phraseology has been introduced, especially in cognitive science, to refer to them—such as "symbol grounding" and "the knowledge level."

It is therefore essential to be clear about one's use of terms. In what follows I will use 'semantics', in general, and 'reference' more particularly, to refer not primarily to computationally- or internet-internal relations, but rather to the relation between language, symbols, semiotically or semantically significant vocabulary or structures, on the one hand, and the (typically distal) objects in the world that, informally, they are "about." Thus in saying "Mother Theresa" I refer *to a person who lived in Calcutta*. What I take to be the primary semantic relation between my present use of "Tutankhamun" and its referent is a relationship that spans thirty-five centuries. And I take the term 'Alpha Centauri' to refer to not to anything on earth, but to an object 41.5 trillion kilometers away. If, therefore, I

[8] «Cite: common logic»

[9] See the sidebar "Classical ontology vs. metaphysical stance" on p. ■■.

type in 'Alpha Centauri' on a web page, and am as a result directed towards, say, the Wikipedia page describing the star, or given a UID, then while referential issues may *warrant* that resulting behaviour, what I am *delivered* (am directed to, see on the screen, etc.) is at best something co-referential with my original term. In all three cases, to put it mildly, the laws of physics prohibit the delivery of the referent itself. (And again, the fact that Alpha Centauri is 41.5 trillion kilometers away is something with which realists, idealists, constructivists, formalists, and intuitionists would all agree.[10])

Needless to say, it does not follow that one cannot refer to items on the web. I have just done so, in fact—in that very sentence. And it may be, in fact undoubtedly is, that there are symbols or structures on the web that refer to other symbols or structures on the web (an IP address, for example, may ultimately be claimed to name a particular web server). The point is only that reference is neither something we can define, nor something we can restrict to a web- or internet-internal relationship—on pain of "disappearing" the structure of mind, sundering the relation between the structure of the internet and the nature of human understanding, and (to put it not neutrally) vitiating our chance of ever developing architectures that truly support human practices. Forewarned is forearmed.

That said, the second class of problems with the current architecture of the web have to do with semantic problems is the sense just adduced. I am not going to argue for a realist approach. On the contrary, as already stated, I will argue that we need to embrace radically constructivist epistemologies in order to provide support for discursive engagement. What I do believe, however, is that until we develop semantical frameworks that are *semantically coherent or intelligible*, in the sense of dealing soundly and genuinely deal with issues of real-world reference (however those issues are named), we will not be able to architect a web that supports those practices appropriately.

The ontological and semantical problems are related. As I will argue, they especially come together around issues of representation. Large swaths of the current intellectual landscape (especially including cultural theory, much of the humanities, feminist epistemology, science and technology studies, etc.) have developed an adversion to representation, to the point that I tend to encounter anti-representationalism embraced with close to ideological fervour. As I have said elsewhere, I am sympathetic to many of the particularities of this rejection—but I take those grounds to provide sufficient grounds only for rejecting classical notions of representation—not rejecting what representation *could be* (in fact, of course, *is*, and "always already" *has been*). That is: one of the tasks that must be taken on, in defending the proposal to be made here, is to resuscitate or reconfig-

[10] Where they would disagree is on what it means to be 41.5 trillion kilometers away, what warrants a truth-claim about the fact that Alpha Centauri is 41.5 trillion kilometers away, etc. In the vast majority of cases, substantive differences among metaphysical positions have not so much to do with *what is true*, and *what is real*, as with *what it is to be true*, and *what it is to be real*.

ure a notion of representation (or something like it) that avoids the perils of classical representation, and thus can serve, in a progressive way, the needs of architectural support for DE.

1d — Plan

The paper is structured as follows. Sections 2–4 presents a number of motivating examples of each of the three types of issue mentioned above—*identity*, *reference*, and *multiple registration*. Following some discussion of how these issues are treated in our current web architectures, Section 5 introduces paired notions of reference and registration, and argues (in their terms) that a constructive, perspectival/contextual theory of identity is needed in order to understand the realm of files, documents, and discursive human practice. In section 6 I argue for the pragmatic possibility of identifying a "level of analysis" of discursive practice above[11] the diverse interchange protocols and document standards and forms of textual structuration currently in use, but safely *below* the level of expressive content itself.[12] In general I will call this the **level of reference** or **referential level**. In section 7, I argue that no existing formal language or calculus (including formal logic, XML, OWL, RDF, data bases, existing knowledge representation schemes, etc.) is adequately structured to cope with the issues of identity required in order to provide internet support for intertextual and inter-discursive reference. Current tools, that is, intrinsically block us from developing an appropriate "ontology" of documents or discursive exchange. Coming out of that analysis, I propose the development of what I call the "**fan calculus**"—a recursive, reflective, descriptive calculus based on a commitment to the sorts of perspectival sense of identity argued for in §3.

Developing the fan calculus will be a major project. My conservative estimate is that producing a usable first version will take from three to five years of dedicated research. Designing architectural support for discursive practice in its terms is a much larger project—though the projects could overlap in mutually constructive ways, and whereas the former will require a focused, dedicated effort, the latter will be more appropriate for collective development, on the model of open-source software development. In sum, I am under no illusion that what is presented here offers a short-term fix for the cacophony of current habits.

At the same time, I believe the time is right for such a project to be enjoined. The advent of mobile computing has caused a juncture in the computational world that allows for fundamental shifts in direction that seemed unthinkable even as recently as ten years ago. Further dislocations will come—for example, in the chance that over the next several decades computation and digital technology will revolutionize the 3D world as much as they have revolutionized the 2D world

[11] Higher or more "abstract" than, one might say—though once we problematize identity sufficiently, the adequacy of a simple "abstraction hierarchy" will begin to crumble.

[12] I take it as a fundamental design principle to avoid even suggesting that we "systematize" the content of human creative expression. In this way the proposal differs profoundly from proposals to build a "semantic web."

in the past. Plus it is hard not to be motivated by the sheer power of human creativity that could be unleashed if we could genuinely deliver on what we have not yet achieved: the web's promise of providing a discontinuously powerful substrate for substantial human engagement. For all these reasons I believe the project, though perhaps foolishly ambitious, is nevertheless worth commitment.

Rethinking the fundamental architecture of the web—and overturning Western metaphysics in the process—is not a task for the faint of heart. But no less is required before the internet can reach its potential of serving as an appropriate substrate for human discursive engagement. Or so I will argue. It is daunting to embark on the project of rethinking things at this level. If we can pull it off, though, history will laugh at the idea that we ever used anything else.

2 · Identity

Documentary identity conditions are stupefyingly complex. Though seemingly innocent, the question of what it is to be one document, as opposed to another—or as opposed to two—can test the bounds of logical analysis. Among other things, as shown by the examples below, the issue is more complex than one of simple "coarse-graining." If we are to support discursive practices with any subtlety at all, we need to recognize that the abstractions involved in establishing documentary identity conditions *cross-cut* in myriad ways.

I will defer questions about how to *treat* these complex identity issues until a later section. Here I want merely to show what they are, in order to convey a sense of what is at stake. The examples are arranged in two groups: (i) human documents—i.e., of a sort that people write, expressed in natural language; and (ii) computational documents, in the sense of being computer-interpretable (possibly computer generated as well, though not necessarily), including programs and files.

2a — Natural language examples

1. **Works:** Complexities impinge even at the level of established "single works." Thus not only can we refer to Kant's *Critique of Pure Reason* as a whole; we can also: (i) distinguish first its first (1781) and second (1787) editions, without regard to translation; or (ii) restrict our attention to its English translation, without regard to the differences between the two editions; or (iii) single out the English version of one or other edition. Similarly, we can further distinguish Norman Kemp Smith' translation from that of Paul Guyer and Allen Wood, or, re the former, talk of its first publication (1929) or its reissue in 1969, or speak of the second printing of that reissue, or of the typographical error in the first print run of that reissue, or identify my particular tattered copy. And so on.

   Many of the issues and considerations that affect whole works have been investigated in the FRBR project ("Functional requirements for bibliographic references"), whose 142-page report has undergone a continuous series of modifications and corrections since its "final" release in 1997.[13] Overall, FRBR distinguishes *works*, *expressions*, *manifestations*, and *items*—but these categories are just the tip of the bibliographic iceberg—and by themselves do not explicate the complex interweaving of issues illustrated above (e.g., of how "translation" and "edition" can cross-cut). And more seriously, the FRBR document itself recognizes that its proposal is at best an informal and initial guide to a domain that may ultimately be of arbitrary complexity.

2. **Time:** Imagine writing on a student paper "this paragraph needs improvement"—or an a later version, "this section is much improved." And consider the two terms "this paragraph" and "this section." They cannot refer to the paragraph as it stands—i.e., to anything associable with the string of characters or words that, for example, would be pasted onto the clipboard if one were to press "copy" in Microsoft Word. *That* paragraph simply is the paragraph that it is—infelicitously composed, we may suppose. Similarly, the ensuing "this section is much improved" cannot refer, as one might say, to the sequence of characters that constitute it, since that text, presumably, has not changed. Rather, what both cases show is that phrases such as "this paragraph" and "this section" must, on pain of incoherence, refer to temporally enduring and in a certain sense abstract but nevertheless temporally-changing entities, of which particular character or word sequences are at best *temporal manifestations*.

   Temporal complexities bedevil the identity of web documents. It has become something of a trope to note, in a URL-based citation, the date on which the page was (last) visited. But whether, if they differ, the Aug. 3 and Sept. 27 renderings of the page pointed to by "the same" URL constitute two different version of the same page, or two different pages, is not an issue that that practice resolves.

3. **Renderings:** On published paper-based books, it is standard to refer to individual words and sentences and such with page references, sometime amplified with references to specific lines and word numbers (such as in "delete the second occurrence of 'decisive' in the first sentence of the third paragraph on p. 73"). The emergence of digital displays, with varying sizes of windows, variances in the use of fonts, the "separation" of content and display encoded in XML and XHTML, etc., have made it obvious that such references are not stable across different renderings of "one and the same document". Needless to say, this causes havoc for those who want such fine-grained reference to document internals across such diversities of presentation (for example, to take an increasingly common example from my own experience, how to refer to passages in scholarly discussion groups when different members, even if all meeting in the same room, are reading the "same" document on different digital devices).

[13] http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf

The rise of PDF to some extent ameliorated some such challenges, which emerged when .doc files started to be widely distributed. But PDF's fixing or "stabilization" of page rendering may more reflect a vestige of the print-based era than a solution to …

4. **Annotations and commentary:** Suppose Hillary sends a draft of a memo to a group of colleagues for comments, and that Trevor, one of the commenters, writes extensive notes on his copy and sends it back. Imagine, too, that three other commenters also make suggestions, so that Hillary has four different "marked-up copies" to respond to.

   It is obvious, first, that although Trevor was making changes *on* his copy (as we would say if that copy were on paper), or *to* his copy (if, as we may suppose, it was a Microsoft Word document), in terms of intent, what it was that Trevor's comments are *about*—i.e., the document they are *targeted towards*—is *not* the copy he has causally affected. Sometimes, it is true, one *does* want to make notes on one's own version—reading notes, for example, such as marginalia, and perhaps annotations from which one plans to speak. But in this case, the document on which Hillary asked for comments is more abstract—something, as we might imagine putting it, that Trevor's copy is a *copy of*. And yet 'copy' is not right either; because what Trevor's copy is copy of, almost certainly is the *version* on Hillary's computer—the one she was working on, just prior to sending out the note. And it is not *that concrete particular* about which Trevor has views. Rather—and this will matter—he is commenting on something *more abstract* than that, more abstract than anything sufficiently concrete to be the result of, or to be the cause of, material or causal changes, but at the same time something *more specific* than the memo that will result from the process—the memo that was improved, en route, by the comments he made on it. And to see the ironies multiply, note that in the sentence "Trevor's comments on this paper vastly improved it," the phrase 'this paper' and the anaphoric reference 'it' do not co-refer (as one can tell from the fact that we might sometimes instead say "Trevor's comments on *an earlier version* of this paper vastly improved it").

The foregoing examples are all aimed at relatively fixed documentary forms—primarily whole books and papers. When the scope is expanded to include email posts, comments on threaded discussion boards, etc., the issues of identity multiply correspondingly (e.g., when one "quotes" a previous post on a forum or email exchange, what is "quoted" is not the same as what is "cut and pasted"—the latter being at least several levels more concrete than the former). But even these few examples will convey a sense of the ontological and identity-based complexities that, impressively, we regularly navigate in the course of our discursive practices.

2b — Computational examples

A similar plethora of cross-cutting individuation criteria applies to computational objects. Web pages "change," corporations distinguish different cultural as well as linguistic versions of "their home page", the "same" web can be hosted on many different computers, etc. And whereas the spray of cascading and cross-cutting forms of documentary identity in the realm of natural language expression are rarely if ever codified, in more computational realms large bodies of practice have developed to manage at least some of the complexity, in at least some corners of the networked world.

Clustered servers are one example: groups of machines assigned to "one" IP address. Replicated data bases are another; complex protocols have been developed to ensure consistency between and among different "instances" of a "single" data base, when causal changes are (inevitably) made to those different concrete instances, but in distributed and even haphazard fashion, but "the data base" that the changes in fact impinge upon is not any one of the instances, but something like a more abstract "regularity" or "invariance" transcending them all. Transaction logs, journalled file systems, etc., go to spectacular lengths to prop up…not the *illusion*, but the *reality*, of a "single" integral unity. Yet another example is provided by the intricate cache coherency protocols employed on multi-processor chip designs, so that a single, coherent "story" about the state of memory is maintained in spite of diverse activities and locations constitutive of it.

Another suite of examples, again backed by substantially sophisticated software systems, is provided in the context of what is known as version control systems for software development—illustrated by such systems as Subversion and GIT. In any situation in which multiple people and even teams are working on a complex software system, different parties can update or revise different copies or instances of the code base, potentially leading to a complex profusion of "copies" of the software in different states. Sometimes different versions are intended. For example one team may start modifying the code for a different platform; another may update the protocol suite on which it is based, etc., without any intent of these modifications being "brought back into" the originating source. In other cases it is intended for the various different changes to be integrated into an emerging common shared base. A complex suite of protocols have been developed, using such terms as 'snapshots', 'clones', 'versions', 'forks', etc., and involving such activities as *checking out*, *signing in*, *committing*, etc. Details don't matter here; the point is only that, in this one (critical) case of managing large software projects, an entire small industry has developed to track and coordinate the complex identity conditions that inevitably arise.

Not all forms of complex file identity are backed with technical standards and systems, however. Imagine a user uttering the following statement about a file:

> "<u>It</u> got corrupted on my hard disk, but fortunately I was able to retrieve <u>it</u> from backup, though I still had some work to do on <u>it</u>, since <u>it</u> had been changed numerous times since then."

At least three different (but obviously intimately related) file identities are implicated by the four underlined uses of the term 'it' in this passage, and perhaps four, depending on one's ontological parse of the situation.

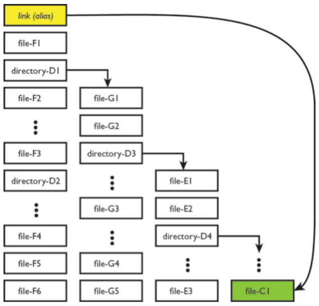


Figure 1

Another case, which arises in the case of backing up and synchronizing file directories, illustrates an as-yet unreconstructed but common situation regarding file identity, mishandled by virtually all backup systems of which I am aware. Suppose a directory is structured along the lines depicted in figure 1. The “link” is a shortcut (variously known as an alias or symbolic or hard link), at in top-level directory, giving you direct access to file C1. which is several levels down in the directory structure (such that accessing it in the normal way would require drilling down through all the intermediate directories). Suppose, in addition, that one wishes to copy or synchronize the whole directory to another computer—say, from an office computer to a laptop one is taking on a trip. The situation one imagines—and expects—is given in figure 2. One runs the backup or synchronization software, and heads off on one’s travels. At some remote location, one clicks on the link indicated at the top-left corner of figure 2b, expecting this to open file C1. Instead, however, the computer responds with an error, or with the message “please mount the disk on your office computer back home.”

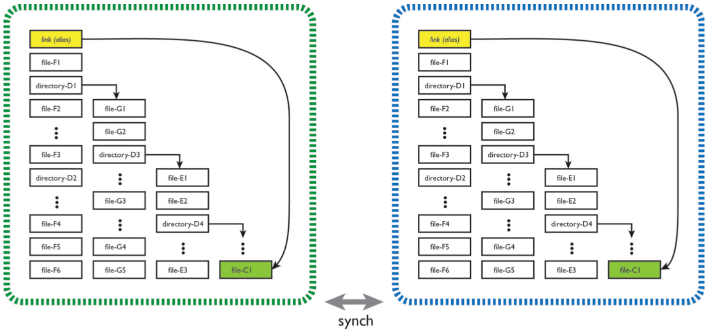


Figure 2

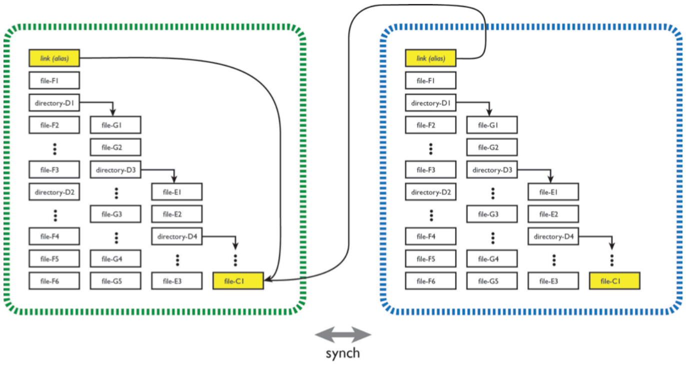


Figure 3

The problem is that, even though the synchronization program copied file C1 from the office computer to the laptop, the *pointer* contained in the link pointed at the *concrete instance of the file on the office computer*. When it copied the files onto the laptop, that is, it created the situation depicted in figure 3. Since one has copied that “instance” of file C1 to the laptop, however—since (ex hypothesi) the synchronization program made that copy, in fact—the “instance of the link” on the laptop should have been updated or “resolved”, one might say, to point to the copy of C1 that now exists on the laptop.

Perhaps that is so. Perhaps the synchronization program should be "fixed" to perform such adjustments automatically (it is exactly this sort of thing that are involved in the replication of data bases). But instead of voicing an opinion on whether that would be preferable, I would make two comments:

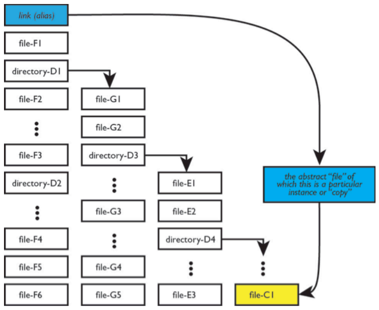


Figure 4

1. There are situations in which one wants exactly what the synchronization program did (i.e., wants the situation depicted in figure 3). A situation might arise in which one *wanted* a link might to *that specific instance of the file* (for example, if it had generated read errors, over the last several times it had been accessed, and one wanted to keep a record of it to ensure that the disk platter it was located on was replaced). In such a case, having the link be automatically "updated" to "follow" file C1 might be exactly contrary to the intentions of the link's creator.

2. More interestingly, in the default case, in which one does want the link updated to point to the copy of the file on the laptop, what is "really going on" may be better described not in terms of *any* specific copy or instance of the file, but in terms of the more abstract unity of which the specific versions are copies and/or instances. As in the case of the natural language documents described above, that is, the "file" in question—the file to which the user wants access—is not the concrete bit patterns on a particular magnetic disk, but a more abstract entity, of which those bit patterns are somewhat contingent realizations. This situation is depicted in Figure 4.

… as far as I have gotten …

3 · Reference

To an extent, relations and dependencies among human expressions have been encoded in the ubiquitous notion of a **link**—particularly the "uniform resource locator," or URL. In fact the fundamental architecture of the internet is based on the idea of a "web of *hyperlinked documents*." In spite of being virtually taken for granted, however, the notion of a link captures only a very small portion of the structure of the interconnectivity of discursive engagement. This is betrayed by two striking facts.

First, other forms of reference remain fully in force—including, to mention just a few salient examples: (i) *citation standards* for books and articles; (ii) *quotation* and other forms of direct inclusion—e.g., in threads, email, forum posts, etc.; (iii) *informal references* such as "your email of Nov. 17, 2009," "invoice #5012," "Derrida's introduction of the term 'différance' in his 1963 paper 'Cogito et histoire de la folie',"[14] etc.; and (iv) pointers, copies, symbolic links, versions, etc. in file systems and software repositories (e.g. soft and hard links in Unix, filepaths such as Users/Ichabod/Documents/Harpsichords/Einstein–32.pdf", "Safari V5.0.3 [6533.19.4]", and so on). Second, and more seriously, the web has failed to unleash a transformation in the forms of interdocumentary citation and reference in wide sectors of society—including, notably, much of academic scholarship, especially in the humanities.

Before exploring what will be required for a web architecture that supports reference, it will be helpful to have a grasp of the range of reference types that will have to be dealt with, an at least an initial sense of the ways in which they are likely to be used.

Table I lists some common varieties with which we are familiar from literary practice prior to the arrival of digital documents. They use a variety of mechanisms for *explicitly* identifying the reference's **target**.

Table I — Explicit forms of literary reference

	Type	**Examples**
1	Citations to *whole works* (as supported by BibT$_E$X, Zotero, EndNote, etc.)	· Derrida, Jacques, 1978. Cogito and the History of Madness, *Writing and Difference*. Trans. A. Bass. London & New York: Routledge · Thesleff, H. Platonic Chronology. *Phronesis* **34** (1989): 1-26.
2	Citations to *regions or passages* within works	· §3 · pp. 127–83 · pp. 49 ff. · op. cit., loc. cit., ibid. · Vol. III
3	Intra-documentary references	· Tables of contents · Footnotes (see Figure 2) · Section references · Indexes

[14] Note that Wikipedia's entry documenting Derrida's introduction of this term contains five non-URL references (and Wikipedia is surely as canonical an instance of a "networked document" as any that exists): two instances of direct quotation, two instances of standard textual citation, and one instance of informal reference ("Schultz and Fried in their vast bibliography of Derrida's work"):

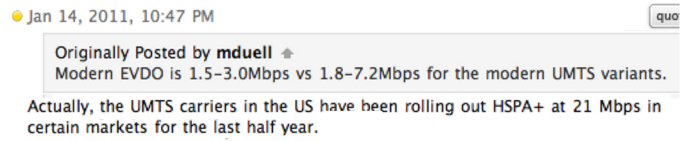
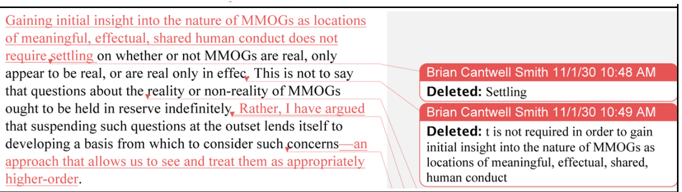
> "'The economy of this writing is a regulated relationship between that which exceeds and the exceeded totality: the différance of the absolute excess.'(Derrida, J., 1978. Cogito and the History of Madness. From *Writing and Difference*. Trans. A. Bass. London & New York: Routledge. p. 75.) Schultz and Fried in their vast bibliography of Derrida's work cite this sentence as where "*JD introduces différance*" for the first time. (Schultz, W.R. & Fried, L.B., 1992. *Jacques Derrida Bibliography*. London & New York: Garland. p. 12.)" [http://en.wikipedia.org/wiki/Différance#cite_note-0; retrieved 11:29:40 am, January 28, 2011]

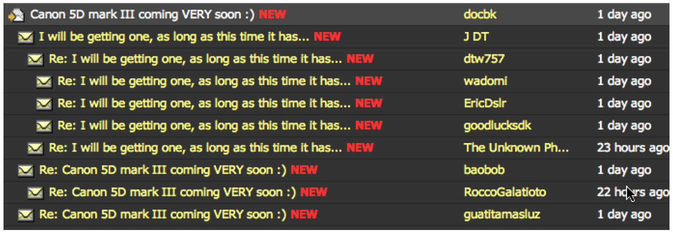
4	Informal discursive reference	· "Your letter of 19 April, 1894" · "Peirce's analysis of *thirdness*, introduced on p. 423 of the *Collected Papers* (Crossfield edition), and elaborated in …" · "That point you keep making" · "The word 'disspirited' on line 3, 2nd ¶, p. 219 which (strangely) should be spelled 'dispirited'." · "What he said" [posted on a forum thread]

As well as reproducing all of these in digital form, we have introduced a number of new explicit referential mechanisms to deal with the web, give in Table II.

Table II — Explicit forms of network reference		
1	Web resources:	· URLs: http://humanities.utoronto.ca/event_details/id=338 · URNs: urn:mpeg:mpeg7:schema:2001 · DOIs: 10.2224/2007-1-29-CENDI-DOI · Email addresses: brian.cantwell.smith@utoronto.ca · FTP file locators: ftp:/ftp-servers.com/mydirectory/myfile.txt
2	Files	· File paths: /Volumes/OS X Server Data/Users/everyperson/ · Version numbers: MSWord Version 14.0.2 (101115)

Along with these explicit forms, a great many referential cases, especially those to the fine-grained internal structure of documents (about which more below) identify their referential targets implicitly or contextually, by exploiting various properties of their concrete physical location, aided by lines, arrows, highlighting, etc. A few familiar examples:

Table III — Implicit forms of reference		
1	*Quotation* in email & posts	 Jan 14, 2011, 10:47 PM Originally Posted by **mduell** Modern EVDO is 1.5-3.0Mbps vs 1.8-7.2Mbps for the modern UMTS variants. Actually, the UMTS carriers in the US have been rolling out HSPA+ at 21 Mbps in certain markets for the last half year.
2	Documentary comparison	 Gaining initial insight into the nature of MMOGs as locations of meaningful, effectual, shared human conduct does not require settling on whether or not MMOGs are real, only appear to be real, or are real only in effec. This is not to say that questions about the reality or non-reality of MMOGs ought to be held in reserve indefinitely. Rather, I have argued that suspending such questions at the outset lends itself to developing a basis from which to consider such concerns—an approach that allows us to see and treat them as appropriately higher-order. Brian Cantwell Smith 11/1/30 10:48 AM **Deleted:** Settling Brian Cantwell Smith 11/1/30 10:49 AM **Deleted:** t is not required in order to gain initial insight into the nature of MMOGs as locations of meaningful, effectual, shared, human conduct
3	Markup, Annotation, and Commentary	· Handwritten commentary on a student paper (figure 3) · Systematized intra-documentary reference (figure 4) · Annotations in PDF and other systems

4	Threads on forums, in email, in annotated documents	 Canon 5D mark III coming VERY soon :) NEW docbk 1 day ago I will be getting one, as long as this time it has... NEW J DT 1 day ago Re: I will be getting one, as long as this time it has... NEW dtw757 1 day ago Re: I will be getting one, as long as this time it has... NEW wadomi 1 day ago Re: I will be getting one, as long as this time it has... NEW EricDslr 1 day ago Re: I will be getting one, as long as this time it has... NEW goodlucksdk 1 day ago Re: I will be getting one, as long as this time it has... NEW The Unknown Ph... 23 hours ago Re: Canon 5D mark III coming VERY soon :) NEW baobob 1 day ago Re: Canon 5D mark III coming VERY soon :) NEW RoccoGalatioto 22 hours ago Re: Canon 5D mark III coming VERY soon :) NEW guatitamasluz 1 day ago See also figure T.
5	*Commentary*—e.g., Talmudic practice.	See figure N.
6	(Annotated) bibliographies	See figures P & Q.
7	Translation	· E.g., recto-verso versions, as in Figure M.
8	Editing	· The "referents" (i.e., entities acted upon) by all commands defined in the object-action" user interface model that is virtually universal in GUIs—thus, all eidting commands in MSWord, InDesign, email clients, etc. …

…

…

3 Uses

Needs to be a section on how references are used—e.g., citations, annotation and commentary, extended dialogues (such as the Leibniz/Clarke correspondence), etc. Also introduce the notion of a "slow conversation"…

4 Properties of Reference

Half a dozen properties of references are crucial—distinguish it from previous forms of net-based linkage.

1. Unlink links, DOIs, etc., references are *not necessarily followed*.

Following a name or link to the document or resource it targets is just one way to use the myriad forms of interconnectivity that knit together the fabric of creative expression. It would be rare for a reader, upon reading a passage that referred to the *late Wittgenstein*, to want to set the original document aside in order to have the full text of the *Philosophical Investigations* presented on screen.[15] Similarly, on encountering a comment that *since George Elliot's Middlemarch, English women writers have largely used their own names*, it would be unusual for the reader to divert their attention to the novel itself.

If we develop a general theory of reference, and implement a network architec-

[15] In some situations it would be more likely for the reference itself to be queried—e.g., in order to determine whether its author intended the reference to include the *Blue Book*.

ture in its terms, we may want to include what is currently the default behaviour for links: an ability to "click on the reference" (or some other such simple action), in order to be taken to—or to have delivered—the document or item that the reference is about. But that is just one possible behaviour that a user might take.

2. References are *themselves part of the content* of a document, and as such need to be humanly understood.

As the foregoing discussion suggests, referential expressions are *themselves part of the content of the documents in which they occur*. Most of the examples cited so far (*an email of such and such a date*, a *novel by a certain novel with a given title*, *page 127 of this or that paper*, etc.) are mundane expressions of ordinary English, framed in terms of concepts and categories that anyone reading the reference will understand (author, title, page number, etc.).

The fact that references *are themselves content* (they don't just point to content) raises some of most challenging hurdles to instrumenting a full architecture to support online reference. Among other things, it implies: (i) that references should be expressed in a way that humans understand—as standard literary references are, such as "*Rethinking Marxism*, Vol. **13**, Nos. 3/4, pp. 70–80," and most URLs and DOIs are not; (ii) that any system that tracks and deals with references must be framed in terms of standard documentary ontologies (of authors, papers, titles, page numbers, publication dates, etc.); and (iii) that references must be treated as *referable-to content in their own right*—thereby facilitating recursive annotation, comments on commentary, etc., all of which are staples of intellectual exchange.[16]

The fact that references are framed in terms of concepts and categories in terms of which documents and targets are identified and found intelligible means that systematizing or formalizing reference, on the internet, will be a task that involves issues of knowledge representation and genuine (system⇒world) semantics, of a sort that will challenge our conceptions of computational architecture.

3. In the general case, references reach *inside documents*, not just to documents as units or wholes.

4. References often target *regions* and/or *extents*, not simply "points" or other unitary objects.

Although many traditional referential forms target whole works—books, papers, web pages, sites, etc.—it is also common for literary references to target extended segments or passages internal to documents, such as "pp. 326–43", "chapters 2-4", or "lines 3–17".[17]

[16] In spite of what is said in the text, it is no intent of this document to suggest that systematization of reference will involve parsing natural language or deciphering arbitrary descriptive references—such as "the best novel in the English language" or "the appallingly derogatory email that I received last week." Identifying appropriate concepts and categories within which system-supported references may be framed will be one of the primary design challenges to be faced.

[17] One is reminded of Descartes' *res extensa*, often translated as "corporeal substance," and

In contrast, URLs, DOIs, and other web-based linkages have been solely formulated in terms of singular objects. This limitation has largely shielded them from the necessity of dealing with document-internal structure. The only common "intra-page" reference is the HTML tag or anchor—a device often pressed into service used when reference to a section of a page is intended (e.g., an anchor to the beginning of a named section on a webpage, with the expectation that the user or reader will recognize that what has been pointed to extends only as far as the next heading at that level).

Some of the challenges of developing an adequate general theory of reference will be: (i) to formulate, explicitly, in a flexible and customizable way, appropriate grammars of the internal structures of referenced documents; and (ii) to allow for reference to extents and regions. One issue that must be faced in dealing with the latter challenge is to accommodate the fact that extended references frequently overlap—as for example in a case when one commenter on a passage highlights several words in a sentence, and another highlights a different set that overlaps with the first.

3 • Reference

1. References target documents and other forms of creative expression at *arbitrary and cross-cutting levels of abstraction.*

One of the deepest profound facts about general reference to documents is that the identity conditions on documents—what it is to be one document, as opposed to another, or as opposed to two—are bewildering complex. The issue is not simply one of "coarse-graining." If we aim to support reference in a sufficiently rich way to support common practice, we have to deal with the fact that the identity conditions on documents are not only abstract, but cross-cut in myriad ways.

Thus not only can we refer to Kant's *Critique of Pure Reason* as a whole; we can also: (i) distinguish first its first (1781) and second (1787) editions, without regard to translation; or (ii) restrict our attention to its English translation, without regard to the differences between the two editions; as well as (iii) singling out the English version of one or other edition. Similarly, we can further distinguish Norman Kemp Smith' translation from that of Paul Guyer and Allen Wood, or, re the former, talk of its first publication (1929) or its reissue in 1969, or speak of the second printing of that reissue, or of the typographical error in the first print run of that reissue, or identify my particular tattered copy. And so on.[18] A similar plethora of cross-cutting individuation criteria applies to computational objects. Web pages "change," corporations distinguish different cultural as well as linguis-

distinguished from *res cogitans*, or mind. Descartes took extension to be the distinguishing characteristic of physical entities or concrete substances.

[18] The 142-page FRBR report ("Functional requirements for bibliographic references"), modified and corrected many times since its "final" 1997 release, distinguishes *works*, *expressions*, *manifestations*, and *items*—but these categories are just the tip of the bibliographic iceberg. See: http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf

tic versions of "their home page", the "same" web can be hosted on many different computers, etc. And as everyone knows, the term 'file' can be used across a spectacularly wide variety of cases—as betrayed in the different referents of each use of 'it' in the following sentence: "*It* got corrupted on my hard disk, but fortunately I was able to retrieve *it* from backup, though I still had some work to do, since it had been changed numerous times since then."

Another limitation of current network links is that, to a first approximation, they refer or point to single internet resource: *a* web page, *a* document in a repository, etc. To some extent, the development of uniform *names* and *identifiers* (URNs and DOIs) has approached the issue of complex documentary identity (e.g., to deal with URL changes, with replicated sites, etc.). Moreover, neither the URN nor DOI framework provides a theory of such cross-cutting abstraction or graining, such that there is a systematic (and humanly comprehensible) way of generating one form of reference from another.

2. *Documents are not their digital representatives*

Perhaps the most challenging fact about treating reference explicitly is that it forces us to confront an issue that, in the world of networked resources, we have to some extent succeeded in avoiding.

It has been common, in practice to date, pretty much to "identify" (i.e., conflate—not make a distinction between): (i) a computational or networked entity such as a file or web page or other internet resource; and (ii) and the document that that resource represents. In some cases, such as the U.S. Declaration of Independence, it is manifest that the "document itself"—in this case, the document that played such a decisive role in human history, whose first sentence is a candidate for being the best-known sentence in the human language—exists at a considerably higher level of abstraction than the numerous copies of it that abound on the web. But even in very modest cases there remains a difference between the document itself and *any* file representation of it. Suppose I post (a .pdf copy of) version 0.6 of this paper on the web, you submit comments, I revise it and post version 0.7, and you note that *the last section is much improved*. The phrase "the last section" refers neither to the last section of version 0.6 (which is what it is—not so good, apparently) nor to the last section of version 0.7, which is similarly what it is, but to a more abstract "section" with a history, which has undergone changes over time.

Once again, one might (vainly, I believe) imagine that one could identity a distinct "item" for registrations of documents are every conceivable or potentially useful level of abstraction. It will be argued below, however, that any such proposal would be hopeless. What we need, instead, is a representational system in which *distinctions between and among appropriate levels of abstraction for the purposes of the reference itself can be flexible and conveniently navigated*.

…

…

Quotes from the DOI

> "The most widely known application of the DOI System is the CrossRef cross-publisher citation linking service which allows a researcher to link from a reference citation directly to the cited content on another publisher's platform, subject to the target publisher's access control practices." [from the DOI Overview]
>
> "A DOI® (Digital Object Identifier) is a name (not a location) for an entity on digital networks." [from the DOI Introductory Overview[19]]

Relation to semantic web [[should this be a sidebar?]]

As subsequent discussion will show, the demands of this reconfiguration are so challenging as to seem overwhelmingly impracticable, if not outright impossible. A common reaction will be that the proposal is unrealistic—that only politically, technically, socially, and/or responsible way forward is to "stay the course" and work with what we have. Three considerations, however, suggest that, on the contrary, we should forge the will to meet the challenge. First, the current eruptive spread of mobile devices represents something of a juncture in global network infrastructure. If there is ever going to be a chance to re-architect the web, this may be it. Second, even if what is proposed here seems a hundred times too difficult to accomplish, it is nevertheless no more than one percent of the task of constructing a genuinely "semantic web." More specifically—and not irrelevantly—all the intellectual and technical issues canvassed below are prerequisites to anything that deserves the latter label. So upgrading from links to reference could be viewed as a first step towards that radically more ambitious task. Third, and ultimately most importantly, I believe the case for reference speaks for itself. It is my conviction that if we have the will to make this change, history will laugh at the idea that we ever used anything else.

5 Systematization of Reference

… Aim of this section is to talk about what it would be **systematize** *reference: provide a humanly-understandable language/set of conventions, "capture" all (reasonable) references, etc. Obviously one can't (and wouldn't want to try to) capture all descriptive referring phrases ("the bunk written by that old so-and-so"). Possibly cite email as an example: we use the categories of "from," "to," "subject", and so on,*

[19] http://www.doi.org/overview/sys_overview_021601.html

perfectly intelligibly.

It will take some discussion to convey a reasonable sense of a potential "pattern of use" that is on the one hand easy and natural enough that it could be expected to be adopted, and yet complete and powerful enough to

Talk about backwards pointing (linkages), and possible uses—such as:

1) "who has commented on this sentence",

2) "what was this word in the original version",

3) "has this paragraph changed over the last 3 edits?"

etc.

6 Requirements

… A discussion of the requirements that will need to be addressed by any actual proposal…

…

…

…

A. Register documents at multiple levels of abstraction
   1. Words, sentences, paragraph, sections, chapters, etc.
   2. Characters, lines, pages, spreads, folios, etc.
   3. The location of all copy-editing marks (use ex. at end of *Merriam Webster*)
   4. All normal text-editing operations (e.g., everything available in MSWord)
   5. All professional text-layout operations (e.g., everything available in InDesign)
   6. EMACS: Some cross-cutting grammars: sentences, lines, words, paragraphs, Lisp expressions, C++ code, etc.
   7. Type, characters, glyphs, ligatures, figures (proportional oldstyle)
   8. Margins, columns, recto and verso pages, folios
   9. Lines, boxes, handles, points, groups, "fill",
   10. Go over the editing and layout operations available in MSWord, InDesign, and identify the ontology in terms of which the editing and layout operations are defined
   11. "White-space editor"—mock up examples

B. Flexibility-customization
   1. Describe reflection in a descriptive/declarative language

C. Social aspects
   1. Talk about developing the grammars (registration schemes) in something like an "open source' community way…

… Plus a zillion other things! …

....

1. *Documents are not their digital representatives*

…

Another limitation of current network links is that, to a first approximation, they refer or point to single internet resource: *a* web page, *a* document in a repository, etc. To some extent, the development of uniform *names* and *identifiers* (URNs and DOIs) has approached the issue of complex documentary identity (e.g., to deal with URL changes, with replicated sites, etc.). Moreover, neither the URN nor DOI framework provides a theory of such cross-cutting abstraction or graining, such that there is a systematic (and humanly comprehensible) way of generating one form of reference from another.

…

Perhaps the most challenging fact about treating reference explicitly is that it forces us to confront an issue that, in the world of networked resources, we have to some extent succeeded in avoiding.

…

Identify document with electronic resource

In practice to date, it has been common to "identify" (i.e., conflate—not make a distinction between): (i) a computational or networked entity such as a file or web page or other internet resource; and (ii) the "content" that that resource represents.

…

…

In some cases, such as the U.S. Declaration of Independence, it is manifest that the "document itself"—in this case, the document that played such a decisive role in human history, whose first sentence is a candidate for being the best-known sentence in the human language—exists at a considerably higher level of abstraction than the numerous copies of it that abound on the web. But even in very modest cases there remains a difference between the document itself and *any* file representation of it. Suppose I post (a .pdf copy of) version 0.6 of this paper on the web, you submit comments, I revise it and post version 0.7, and you note that *the last section is much improved*. The phrase "the last section" refers neither to the last section of version 0.6 (which is what it is—not so good, apparently) nor to the last section of version 0.7, which is similarly what it is, but to a more abstract "section" with a history, which has undergone changes over time.

…

…

Ontology document with electronic resource

Once again, one might (vainly, I believe) imagine that one could identity a distinct "item" for registrations of documents are every conceivable or potentially useful level of abstraction. It will be argued below, however, that any such proposal would be hopeless. What we need, instead, is a representational system in which *distinctions between and among appropriate levels of abstraction for the purposes of the reference itself can be flexible and conveniently navigated.*

…

…

Quotes from the DOI

> "The most widely known application of the DOI System is the CrossRef cross-publisher citation linking service which allows a researcher to **link** from a reference citation **directly to the cited content** on another publisher's platform, subject to the target publisher's access control practices." [from the DOI Overview]
>
> "A DOI® (Digital Object Identifier) is a **name** (not a location) for **an entity on digital networks.**" [from the DOI Introductory Overview[20]]

…

[20] http://www.doi.org/overview/sys_overview_021601.html