

University of Nebraska - Lincoln

## DigitalCommons@University of Nebraska - Lincoln

---

Faculty Publications -- Department of English

English, Department of

---

Spring 3-18-2013

### A Matter of Scale

Matthew L. Jockers

*University of Nebraska-Lincoln*, [matthew.jockers@wsu.edu](mailto:matthew.jockers@wsu.edu)

Julia Flanders

*Brown University*, [Julia\\_Flanders@Brown.edu](mailto:Julia_Flanders@Brown.edu)

Follow this and additional works at: <https://digitalcommons.unl.edu/englishfacpubs>



Part of the [Arts and Humanities Commons](#)

---

Jockers, Matthew L. and Flanders, Julia, "A Matter of Scale" (2013). *Faculty Publications -- Department of English*. 106.

<https://digitalcommons.unl.edu/englishfacpubs/106>

This Article is brought to you for free and open access by the English, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Publications -- Department of English by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

*Authors' Note: Copied below are the slides and script of the keynote lecture presented during the Boston Area Days of Digital Humanities Conference at Northeastern University on March 18, 2013. The keynote was a staged debate between Julia Flanders and Matthew Jockers addressing the "matter of scale" in digital humanities research.*

**A Matter of Scale**

*"If our interest in literary study still concerns individual works of literature (however we may define that term), at some point we need to turn our interpretive attention back to these."*

Julia Flanders  
Brown University  
Julia\_Flanders@Brown.edu  
@julia\_flanders

*"Close reading is not only impractical as a means of evidence gathering in the digital library, but big data render it totally inappropriate as a method of studying literary history."*

Matthew Jockers  
University of Nebraska, Lincoln  
mjockers@unl.edu  
@mjockers

JOCKERS:

I'd like to begin by thanking Ryan Cordell and the rest of the organizers of this exciting Boston area Day of DH event.

FLANDERS:

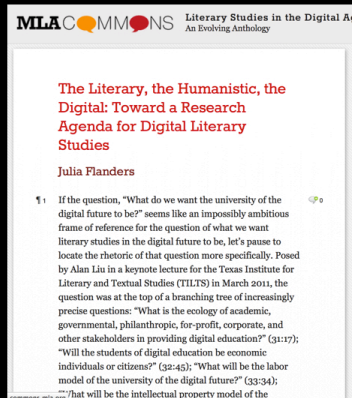
Yes, this springtime of DH feels like an exciting beginning on many fronts. I'm really happy to be here and grateful as well to all who have attended!

In which they discover a problem. . .

2

JOCKERS:

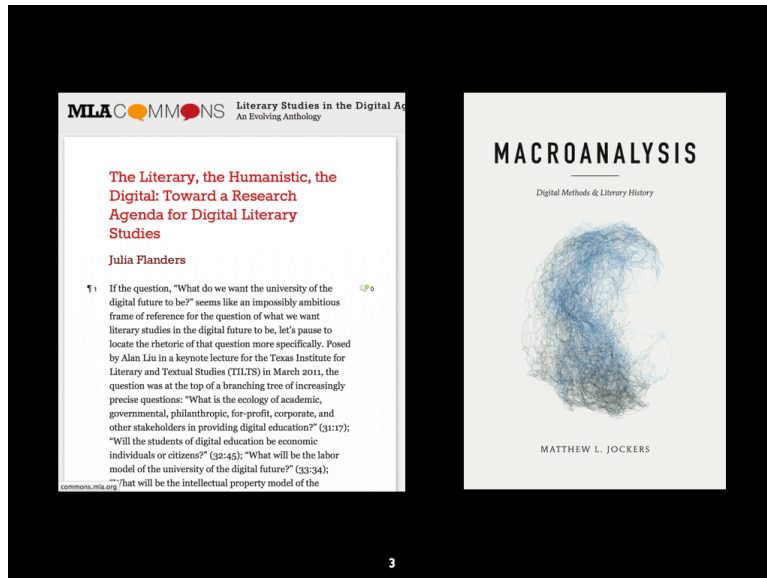
So, Julia, it's really interesting to be standing here talking to you about scale . . . I was deeply influenced by your 2005 article about the importance of detail and how computational approaches to the study of texts can help bring missed detail into our field of view. And now I see that you're back to writing about scale and detail, but now more broadly and more in the context of both "close" and "distant" approaches.



3

FLANDERS:

Yes, Matt, and it looks as if you have too—in fact I've been enjoying your book a great deal, it's given me a lot to think about. I had sort of expected it to be a paean to "big data" but it seems that the situation is more complex than that.



JOCKERS:

Well, yes, I think it is more complex. This stereotype of “big data” is getting so tiresome! ... and, really Julia, I’m a bit tired of it myself. That is, I’m tired of the big data vs. small data battlefield idea; I’m not tired of BIG DATA!

FLANDERS:

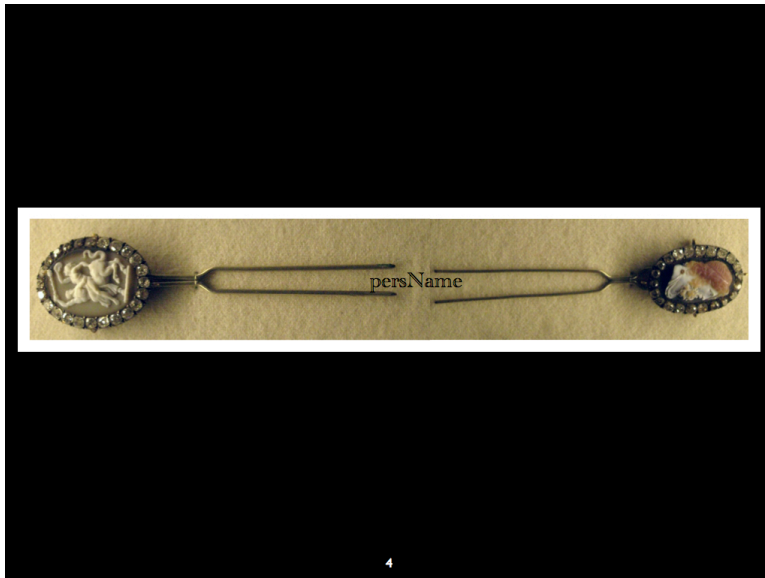
Yes, I agree, there seems to be a lot of talk. Some of it is probably good, but there also seems to be a fair amount of acrimonious, sky is falling type discussion.

JOCKERS:

Right, for some commentators, it is as if these different scales of evidence were somehow mutually exclusive. In my book I write about the bellicose language that has evolved around this discussion. Enough already! Having said that, I don’t think the arguments are all spurious, and if we can cut through the knee-jerk stuff that plays well in the popular press, then I think we find that there are some important points to highlight.

But look, as long as we are confessing, I ought to admit, that when I began reading your MLA Commons piece I assumed you’d be coming out in favor of some sort of organic, locally-grown, back-to-nature theory of craft encoding.





The truth, however, seems to be that you're interested in the same thing I am, which is to say, how "macro" and "micro" approaches are interconnected and interdependent.

FLANDERS [surprised but pleased]

Yes, absolutely. But wait--this is really awkward--we've been invited here for a debate--we can't just agree right away. You were supposed to deliver a knockout blow to the hand-carved artisanal TEI element, and maybe I'm supposed to demonstrate the intellectual bankruptcy of monstrous industrial-grade data.

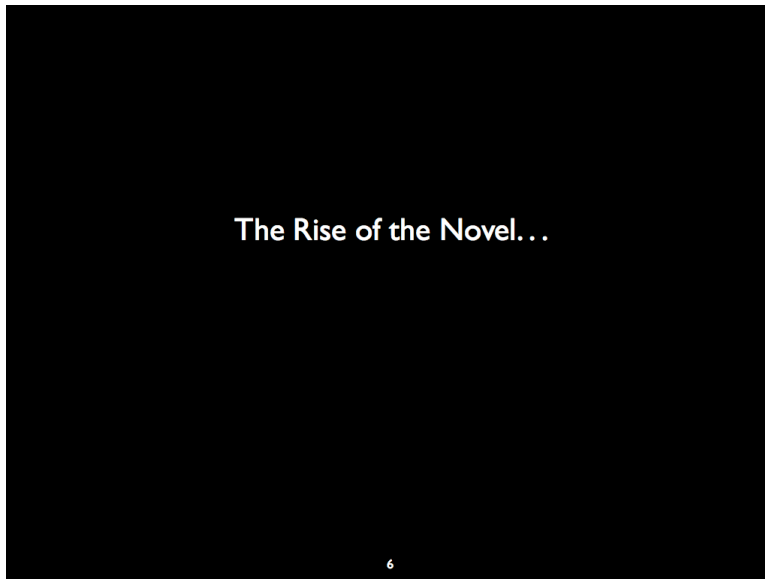
They might not pay our airfare if we don't put on a good show. Let's step back for a moment and at least consider the possibility that there's something to disagree about. Look—how about I defend "small data" and then you defend "big data" and we see what that looks like?

**In which they wax disputatious and  
deal mighty blows. . .**

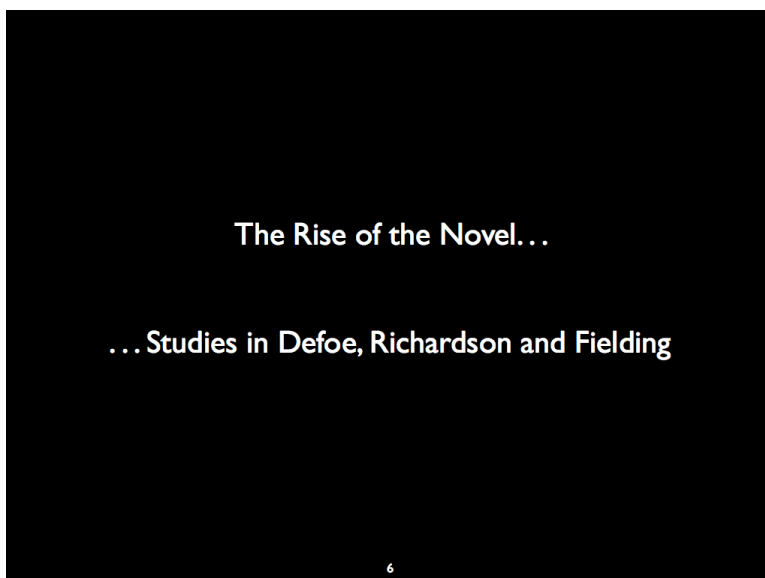
JOCKERS [getting into the spirit of it]

OK, Julia, that sounds like a great idea. And in the spirit of putting on a good show, let me fire the first shot by saying that your super rich, super encoded, super-duper, doubled keyed small data is ultimately anecdotal and arbitrary.

My big data's got context on a grand scale! Access to and analysis of "big data" (and what I really mean here is "big literary corpora") provides us with unprecedented access to the literary record. Consider Alan Watt's magisterial study of the English novel:



*The Rise of the Novel*. It is a brilliant, insightful synthesis of literary history brought to you by one of the great synthesizing minds of our generation.



Now consider the subtitle of Watt's book: "*Studies in Defoe, Richardson, and Fielding.*"

FLANDERS [interrupting]:

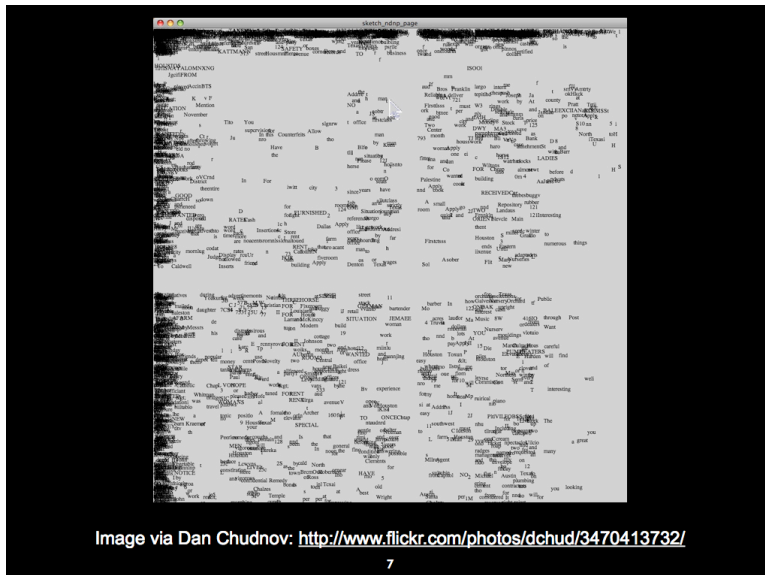
Yeah, I can see where you're going with this—

JOCKERS:

Right, and just as we would not expect an economist to generate sound theories about the economy by studying one or two consumers or one or two businesses, we should not expect sound theories about literature, or at least about literature in the aggregate or about literary history, to be generated out of a study of a few books, even if those books are claimed to be exemplary or representative. Now, of course, it is entirely possible that a study of a few texts might lead us to a very good theory of literary history. But without a context in which to evaluate that theory, we'd really have no way of knowing that it was spot on. We might say, "well, yes, that is well argued and it seems to make sense given what I know about literature and etc. . . but that's about it."

FLANDERS:

Ok, yes, but . . .



. . . isn't that big data of yours full of messy OCR errors and deeply lacking in terms of metadata? I think the classic case for the "micro" approach says, in effect, that we can't trust big data because it's fundamentally careless from a data capture standpoint: it's an industrial product with very little quality assurance, not "scholarly" quality.

JOCKERS [interrupting]:

Big data isn't perfect, so let's all sit on our hands.

## FLANDERS:

Well, maybe we could restate that as “Big data isn’t perfect, so let’s use it in cases where perfection doesn’t matter.” There are going to be cases where a useful research outcome depends on greater precision. But another argument, and probably a more compelling one is that. . .

- we can’t learn much from big data because it’s unstructured
- it’s not self-aware
- it’s just a string of characters

8

- We can’t learn much from big data because it’s unstructured
- it’s not self-aware
- it’s just a string of characters (that don’t necessarily match the characters that were there in the source).

People who distrust big data are often coming from a background where they’re used to working with data that “knows” a lot more about itself, its structure, its contents. I’m very much in favor of large collections, but I find it very frustrating to work with data that seems to say so little about the things I’m really interested in as a researcher. If I’m studying a collection that contains drama, novels, poetry, and letters, it seems obvious to me that those distinctions should be accessible to me in my analysis. Similarly I should be able to reliably exclude from analysis things like annotations, headings, editorial notes, etc. if they aren’t part of the linguistic information I’m interested in.

These kinds of distinctions seem to me to be very much in line with the goals of studying big data--in fact they really help fulfill those goals to a much fuller extent. Without that markup, I don’t see how we can really make interesting and nuanced arguments about literary and cultural texts. Sure, we can infer (sometimes) the presence of these structures from epiphenomena in the text, but not reliably and not always.

## JOCKERS:

Yes, yes! But as much as I agree with you about how lovely that all sounds, that level of markup and detail just isn't practical at the scale of big data!

FLANDERS:

So, I don't disagree with the principle you're stating here, but I don't think it's the most useful way of stating it. How about if we say instead

- "It's a waste of money representing **gratuitous** levels of detail"

"It's a waste of money representing **gratuitous** levels of detail" (i.e. beyond what is needed for the research outcome) and also

- "It's a waste of money representing **gratuitous** levels of detail"
- "Questions that require both scale and detail are going to be more expensive to address"

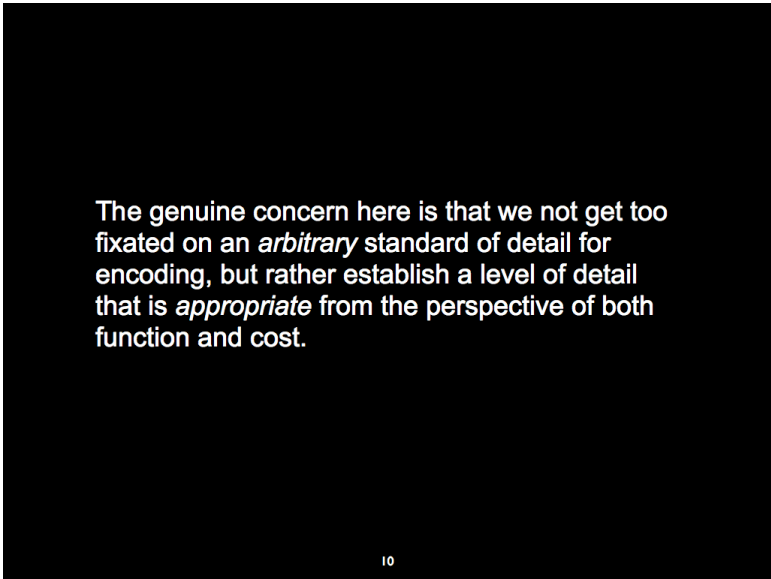
"questions that require both scale and detail are going to be more expensive to address" (but not necessarily too expensive, if they're also of very high value).

JOCKERS:

Yes, of course, I do agree, and I suppose it's worth mentioning that in my own work, I don't begin with un-encoded plain text. Instead I go for a low-hanging markup approach. All my documents have at the least a TEI header that I pack with as much information as I can, and then within the main text, I require structural markup at least down to the paragraph level. I suppose my genuine concern here is that we not get too fixated on perfection in encoding or even in our OCR. We need to accept that, at least for now, we can't have all our encoding and mine it too.

FLANDERS:

And I think I agree; maybe I could restate what you just said as



The genuine concern here is that we not get too fixated on an *arbitrary* standard of detail for encoding, but rather establish a level of detail that is *appropriate* from the perspective of both function and cost.

“My genuine concern here is that we not get too fixated on an arbitrary standard of detail for encoding, but rather establish a level of detail that is appropriate from the perspective of both function and cost.” In other words, let's not treat encoding as an abstract virtue (more is better) but as a strategic tool among others.

JOCKERS:

OK, then we can agree that some level of markup and detail is necessary and that some level is not practical, but what about the larger argument about big data as a way of providing context. After all, I think I may have just insulted Ian Watt.

FLANDERS:

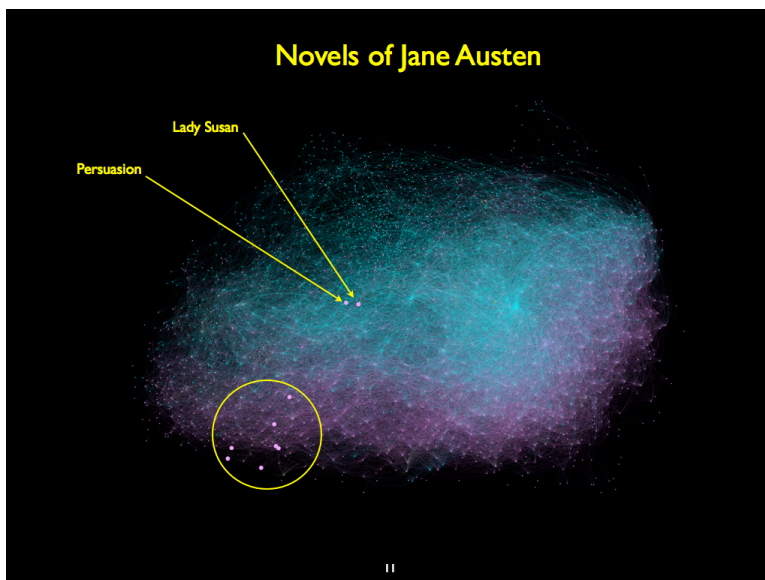
If I get your point, then I think that the kind of broader literary context you're describing would not enable us to either strengthen or refute Ian Watt's book. Rather it asks us to write a completely different kind of literary history. Watt's assertion, I think, would be that the authors he doesn't discuss would not change his argument at all, since his argument has precisely to do

with the assertion that the tradition of “the novel” as we know it consists of a developmental trajectory defined by the authors he does discuss. In other words, it’s their theory of the novel he’s interested in. So our correction of Watt isn’t to say “Hey, Ian, you make a plausible point but we can’t know whether you’re right until we look at a broader sample”—rather we want to say “Hey, Ian, we are not interested in reading literary history that produces a narrative based on what cultural actors say about their work; we are interested in reading literary history that is grounded in a broad data-gathering operation...”

JOCKERS:

Yes, precisely—I suppose what I’m suggesting with macroanalysis is a different kind of argument-making or hypothesis-making. The promise of big literary data is that it can expand the context in which we read Austen and Melville and help us understand how those writers exist inside a much larger literary economy or, if you prefer a more naturalistic metaphor, a larger literary ecosystem.

Honestly, I don’t think I really understood Austen until I saw her in relationship to the other 1800 authors in my corpus. Or maybe understood isn’t the right word. Appreciated is really a better choice here.



I suppose this is as good a place as ever to confess that I’m not a Janeite. I don’t like her novels very much, but I’ve ended up writing about them a great deal because Austen’s books turn out to be quite interesting when seen in context of 3450 other books. And Julia, that’s all I want to say about that.

In which all is forgiven; they gaze upon the  
Micro and the Macro and find it good. . .

12

FLANDERS:

Well, Matt, this has been interesting, and your point about how you came to appreciate Austen seems to make a good case for the macroscale, but I think it's also clear from what you've just said that Austen may also be appreciated in isolation.

JOCKERS:

Exactly, and that just strengthens my sense that these two "sides" are really caricatures. When we look more carefully, "close" and "distant" approaches, micro and macro analysis, are not in conflict or even contradictory.

FLANDERS:

In fact, a similar point was made just a year ago by Alan Liu . . .



Alan Liu @AADH



13



in his keynote lecture at the Australasian digital humanities conference: Liu argued that these two methods we seem so intent on discussing in terms of opposition are not really that different. Liu goes so far as to call close reading a "phantom term." He says that the actual analysis conducted by a close reader is only felt to be close because it focuses on a single poem or passage. There is nothing about the approach, which is to say "the method," that is inherently tied or restricted to individual works of literature.

#### JOCKERS:

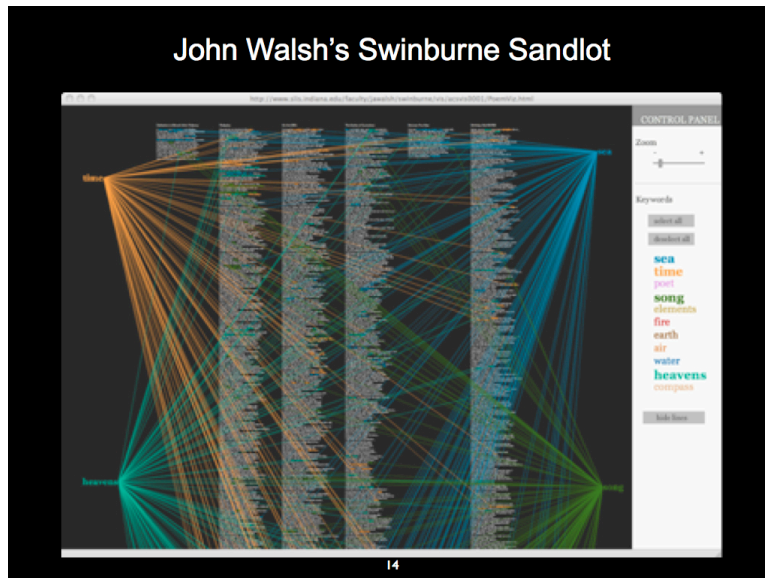
You know, I think Alan would probably agree with me in thinking that close reading is an unfortunate name for a critical practice that preferences the analysis of a certain type of literary data and a certain set of methodological boundaries about what can be said and done with that data. The critical work, the method, is not necessarily or essentially "close." It seems possible to me that this approach of ours has only come to be called "close" because without computing, "close" was all that was humanly possible. If we look under the hood, as Alan has done, we find a methodology that is at its core interested in the careful and sustained explication of detail.

And this business of detail, puts us squarely back onto your familiar turf. Julia, you've trained generations of text encoders to read, to recognize, and to mark up details held within individual texts. Isn't the text encoding process a highly specialized form of close reading? A method that is explicitly designed to anticipate scholarly inquiries of the future and implicitly an act of interpretation.

#### FLANDERS:

Yes, some kinds of text encoding approaches are conducted very much in this spirit. I do think that the *process* of text encoding is very much like a close reading (being both detailed and interpretive), but the *resulting data* can go either way.

Some projects do really focus in on individual texts with the goal of elucidating what's going on—thematically, rhetorically, etc.— so that the patterning of the text itself can be represented; a great example is John Walsh's Swinburne Sandlot:



Digital scholarly editions similarly focus on the interior ecology of a single text.

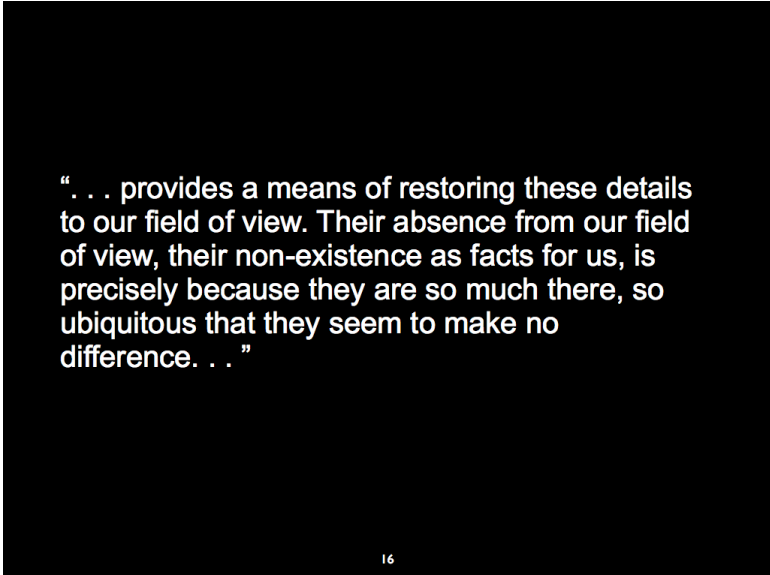
But at the WWP, the detailed markup we're doing isn't aimed at the individual text: I would call it "detailed data at scale"—we are encoding a collection of texts in a consistent way that captures a set of repeating features: e.g. named entities, rhetorical structures, intertextual references, etc. These features operate meaningfully both within the ecology of a single text, and also within the ecology of the collection as a whole: so the "micro" view represented by the markup is very much in the service of the "macro" view represented by the collection and the collection-level tools/interface.

So is this perhaps an example of what Alan Liu was getting at?

In which they deploy their spades and dig deeper  
into the mystery; mighty words are uttered and the  
obligatory word cloud makes its appearance. . .

JOCKERS:

Maybe, yes. I think about it this way: when we explicate a stanza of poem or a passage of text, we engage in a certain type of lens focusing. For me a very good expression of this idea—within a strictly digital humanities context—is found in your own 2005 article on this subject. The article titled “Detailism, Digital Texts, and the Problem of Pedantry” discusses how the great stylometrician John Burrows’s uses computation as a way to bring the most common words in Jane Austen’s novels, words such as “the and of,” into our field of view. In that essay, you write about how . . .



“. . . provides a means of restoring these details to our field of view. Their absence from our field of view, their non-existence as facts for us, is precisely because they are so much there, so ubiquitous that they seem to make no difference. . . .”

16

the computation “provides a means of restoring these details to our field of view. Their absence from our field of view, their non-existence as facts for us, is precisely because they are so much there, so ubiquitous that they seem to make no difference” (2005, 56–57).

I think it is a rather easy step from that comment of yours about close text analysis to what I have come to call macroanalysis. The objective is much the same: to restore to our field of view precisely that which is right beneath our nose but too ubiquitous to be synthesized in the human mind.

FLANDERS:

So, in essence, making it possible to see both scale and detail simultaneously?

JOCKERS:

Yes.

FLANDERS:

So, in terms of the actual object of analysis, you’d say this matter of scale is, in a strong sense irrelevant. Whether we are explicating the details of a poem or of a genre or period, we are still — necessarily — studying some subset of the whole. To say that the macro scale ignores the

nuances of the individual text is a specious argument. All analysis ignores one nuance or another.

But maybe it doesn't need to! I really think that there are research questions that require an analysis that can take advantage of both.

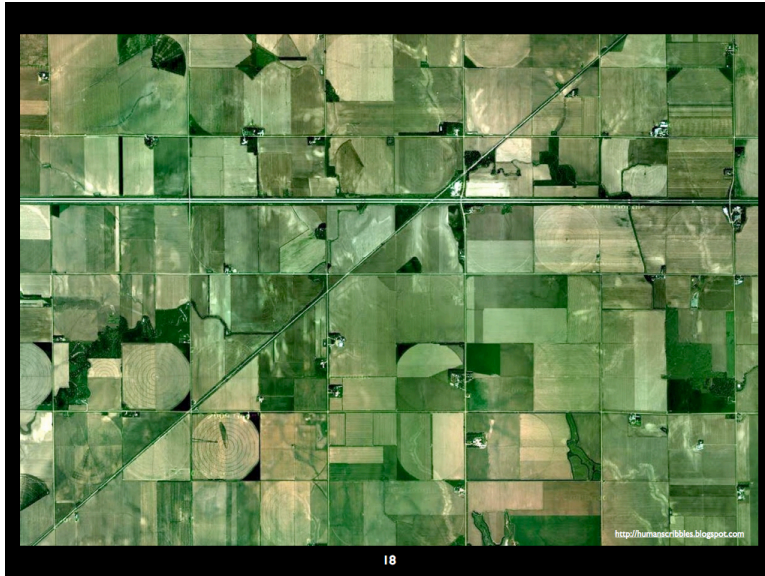
JOCKERS:



Geez, Julia, it sounds like you're talking about some sort of harmonious synergy ;-). But let's see if we can figure out how it would look to do both . . . Let's start with an analogy:



When you drive through the fine state of Nebraska, you see the individual rows of corn, the silos, the barns, and so on. These are the details you see from up close. You don't see how each farm is connected to another farm in a beautiful and organized patchwork of 640-acre sections.



You don't see how the landscape of one farm is the result of and dependent upon the landscape of those surrounding it. You need a plane or a satellite to reveal those particulars.

FLANDERS:

So, close or distant reading, whether done with the assistance of the computer or with the naked eye, is a method dependent upon different levels of focus and attention; at one moment we focus our critical lens in a way that is meant to call certain aspects of that poem or passage into our field of view and in that moment we necessarily ignore other facets that might be seen using a different focus.

JOCKERS:

What is fascinating for me is the way that computation can be leveraged in this process. In your article you write specifically of Burrows's use of the computer to help him see more in the texts that he was then reading or studying. The further step, beyond Burrows, is to allow the computer to help us go even deeper, to go beyond what we are capable of seeing at the level of an individual text. I'm reminded here of a point made by Tim Lenoir, the historian of science. Riffing on Ian Hacking's argument that electrons are real when you can spray them, Lenoir, suggests that quarks would not exist were it not for the particle accelerators that were built to discover or produce them.

FLANDERS:

In other words, our tools bring the data into existence, not just into view? That's a remarkable and provocative statement, Matt.

JOCKERS:

Well, thank you, Julia. That's how I think about it anyhow. But I think you and I must begin from this point of agreement and now work our way towards what the photographers call depth of field. How do we alter the f-stop and shutter speed so as to keep as much in focus as we can? It is entirely possible that the extraordinary things we think we know about Melville's use of whaling as a theme, or Hemingway's minimalist turn of phrase are not extraordinary at all.

At the micro scale we are interested in and bring our focus to the exceptional or "extraordinary": to the great poem of Milton; to the exceptional novel of Joyce; to the timeless play of Shakespeare. These are, it seems, extraordinary works worthy of close analysis. Conventional wisdom would have us believe that at the macro scale, our focus must necessarily shift away from the extraordinary towards the ordinary; or as Moretti has called it, toward the "great unread," but I think what we are discovering here today is that this really isn't or doesn't have to be case at all.

FLANDERS:

Fascinating--I guess, if we're using a metaphor like "depth of field" here, it's theoretically possible to focus on all planes at once, given the analytical equivalent of a very narrow aperture...but that's a digression.

I would say one of the things a mid-sized collection like the WWP tries to do is identify a scale at which we're not looking at these exceptional texts, or at least not treating them as exceptional or timeless: we're trying to capture enough material that any given text can be understood as similar to or related to (on all sorts of different axes) many other texts: by genre, by time period, by topic.

So the collection doesn't single out individual texts: on the contrary, it embeds them in a web of commonalities and interconnections. But once you notice a specific text for some reason, you can examine it closely and discover what makes it distinctive. We've tried to create an interface that supports that kind of shuttling between different levels of scale: seeing patterns, seeing outliers, zooming in and zooming out. The tools aren't very good yet, but they're getting better.

JOCKERS:

That's right, Julia. I love what you just said about "examining a work closely and discovering what makes it distinctive." That is precisely what I am talking about here. I don't think we can understand what makes a work distinctive without a very large context in which to make that observation. That's kind of like what I was saying earlier about my experience with Austen.

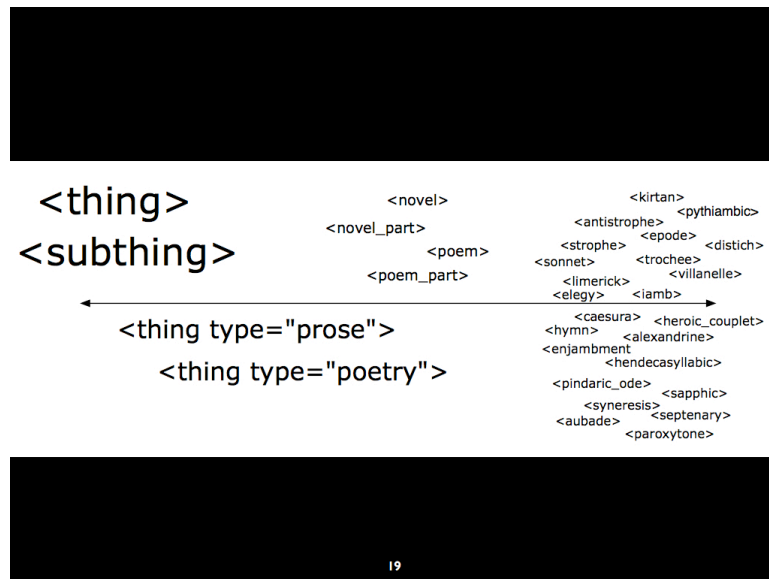
But, you know, I also think there is something more at stake here when we talk about "exceptional" and "extraordinary" and "distinctive." Maybe another way of coming at this is to ask what it is we do about the kind of exceptions that don't lend themselves to neat encoding and computational extraction? I found my appreciation for Austen by studying her stylistic and thematic similarity to several thousand other 19th century authors. But what do we do, for

example, about plot movement and character types? These are not the kinds of things easily detected by my algorithms. How would these things get encoded or extracted by machines?

In your keynote at last year's AADH conference, you talked about specific textual features that are difficult to encode precisely because they're exceptional; they can't be assimilated to a standard representational strategy.

FLANDERS:

Or at least the friction they produce reminds us that there's **representation and strategy going on**. In an important sense, exceptionality is an artifact of the granularity of our representational system. In the TEI, we're always trying to find a balance between doing too good a job of expressing exceptionality, and doing too good a job of assimilating exceptions to a general system. But either way, the "exceptionality" or unrepresentability of a given thing is determined by the information model, not inherently by the thing itself.



So, we seem to have come full circle. We're in agreement that both exceptionality and pattern are interesting and also interconnected — and that micro and macro approaches are really two faces of the same thing.

JOCKERS:

Well, yes, two faces, but how do we get that longed for synergy, that full depth of field? I feel as if I've been able to discover a lot of things with my existing tools and corpus: things having to do with stylistic and thematic change over the course of the 19th century. But one area where I'd love to get some data but can't yet because I haven't figured out how to do so with my existing tools/corpus relates to questions I have about plot. And here I'm thinking specifically about Vladimir Propp and the whole business of archetypal plot structures. I'd really love to tackle plot at the macroscale, but what are the "features" that constitute something like rising conflict or

coming of age. It is an interesting problem that feels unsolvable, but, of course, that is also what makes it worth tackling.

FLANDERS:

I wonder whether your archetypal plot types (a la Propp) might turn out to be correlated with patterns of textual features such as dialogue (length of typical utterance, number of speakers involved in a conversation), tendency to quote other texts, patterns of place names, use of epistolary forms, and that sort of thing?

JOCKERS:

Well, we know that some of these features you mention are correlated to genre. We conducted some experiments in genre classification at Stanford that allowed us to detect Gothic novels and Bildungsroman novels with a pretty good rate of accuracy. But here I'm interested in going deeper than book level labeling: that is, I want to probe deeper and be able to say something more than: yes, *David Copperfield* is a bildungsroman because my computer tells me so. I mean I want to be able to track the actual movement and structure of the plot. If we can figure that out, then we'll be able to see if there are patterns of expression across time, across cultures, etc.

FLANDERS:

This is actually quite exciting. I'm starting to see possibilities for some very provocative and fruitful research based on this combined approach. I've always wished someone would create a text analysis tool that takes markup into account. . . There was a great session at DH2012 on text analysis and text encoding that really got me thinking.

JOCKERS:

Hmm. . . so what kinds of research questions would this system be asking?

FLANDERS:

Well, for instance, what if we do our standard types of vocabulary analysis in novels, you know, word frequency distributions and all that. But what if along side that data we could tell what voice the words were in: the background narrator, the various characters? What if we had data about the gender or social class of the speaker? That's the kind of thing markup can tell us. Wouldn't that let us add more nuance to arguments about, for instance, authorial influence? What if it turned out that Austen had more influence on later styles of character dialogue than she did on styles of background narration?

JOCKERS:

That would be a fascinating thing to investigate, Julia. What you are describing here is text-mining nirvana. I have dreams of such utopias.

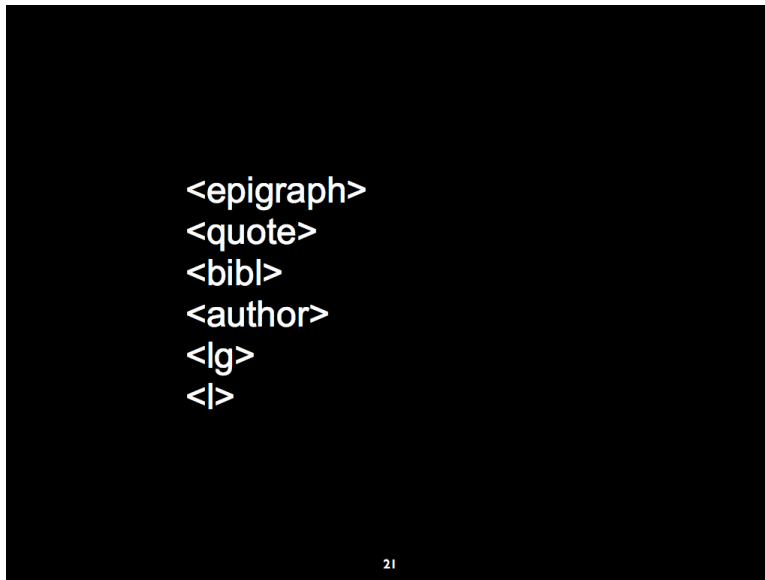
FLANDERS:





FLANDERS:

The TEI gets a lot of bad press because of its flexibility: the idea that you can encode the same thing in hundreds of different ways. This is true, but it's important to think carefully about what that means. In TEI, it's true that there's a daunting amount of structural flexibility: there are hundreds of ways you can combine elements like . . .



<epigraph>, <quote>, <bibl>, <author>, <lg>, <l> to represent the fact that a long poem starts with an epigraph quoting a poem by a different author. But when you get right down to it, no matter how you combine those elements, they are each telling you something very determinate and unambiguous that you would have a hard time knowing otherwise.

JOCKERS:

You've reminded me again that big data is only as good as its big metadata. I love Google's ngram viewer, but as a window into culture and literature, it's a rather dark glass to look through. But tell me why do you think this list of TEI tags might provide information I could leverage at scale.

FLANDERS:

If we focus on what markup can tell us about *what things are* and can also draw intelligent inferences from *the way they fit together*, we can learn a great deal regardless of minor structural variations. Let's take the tags I just mentioned:

Things this markup tells us no matter what:

- There is an epigraph associated with this poem
- The epigraph consists of a quotation
- The quoted text is poetry
- The quoted text is being attributed to a source, about which some information is given
- The poetry consists of a sequence of individual lines
- Each line is poetry: hence may have regular scansion, rhyme, and other properties that distinguish it from prose

22

- there's an epigraph associated with this poem
- the epigraph consists of a quotation that contains some poetry
- an author is being associated with that poetry
- the poetry consists of a sequence of individual lines (each of which can be assumed to have regular scansion, perhaps rhyme and other properties which distinguish it from prose).

In this sense, the real power of the TEI (or of any other XML markup) for text analysis is its ability to localize our insight by giving specific names to the distinctive sites in the text: direct speech, notes, poetry, prose, names, intertextual quotations, apparatus of various kinds. The level of variation in how these things are identified and structured has a comparatively minor impact on the information we can get from them. If I have a corpus that includes some texts that use <name> and others that use <persName>, I still know they're names. If my corpus includes texts that use <lg> in different ways, or omit it altogether, I can still tell poetry from prose.

Anyway, that's my hobbyhorse! But it's exciting to think of the avenues of research this kind of approach could open up.

In which they contemplate difficulties  
and become anxious. . .

23

JOCKERS:

Exciting, indeed! And you can expect a call from me next Monday. . . But you know, it occurs to me that you and I have been drinking out of the same kool aid firehose for a good number of years. It might be worthwhile to pause here and acknowledge a few of the real challenges associated with this kind of work. I worry a lot, for example, about how even our big data corpora are still really small, at least when it comes to making claims about “Literature” with a capital “L.”

FLANDERS:

One thing we wrestle with at the WWP is the problem of what our collection really represents. Back when the project was first envisioned, we thought that we could actually capture all of the extant women’s writing in English before 1830, so representativeness wasn’t so much of a problem. But (I guess we should be glad) that turned out to be wildly wrong—there were orders of magnitude more eligible texts than we had imagined, far more than we’ll ever likely capture before the heat death of the universe at the rate we’re going.

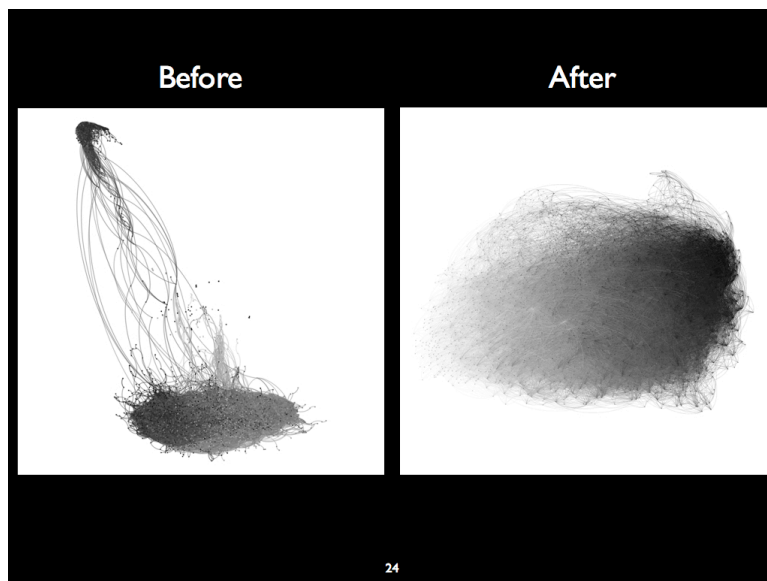
So now, when we offer tools for text analysis that operate on the whole collection, we have the question of what this collection can actually tell us: about genre, about authorship, about periodization, about anything. It’s a mid-size collection, about 350 texts from a wide range of genres, topics, periods, etc., and clearly there’s some very useful information to be gained from studying it, but precisely what kinds of conclusions can one draw? I like very much Steve Ramsay’s idea that the point of such tools is to permit exploration, to pique our interest and prompt further discovery, but if we were to provide tools for statistical analysis, I think they could easily be misleading given the nature of the sample.

That said, I think representativeness is a very vexed question for any collection—even if one is acutely aware of the problem, as the corpus linguists are, it seems that the best one can do is be very, very transparent about one’s collection development strategy, and hope that the user

reads the documentation. But both of these conditions seem fragile... and as text analysis tools become more novice-friendly, I think they're more likely to be used in a novice way. So how do you handle this?

#### JOCKERS:

At some point during my work on the 19th century novel, I had to make a decision to quit collecting texts and start analyzing them. How I got to that point is another matter, but when I began the project I had 950 books and when I made that decision to quit collecting I had 4,700 books. I mined that data and I wrote the last two chapters of my book. About the time I was getting ready to submit the final manuscript, I discovered that there were not 4,700 books. There were actually 3,346. It turned out that the materials my colleagues and I had collected included many multi-volume novels that had not been stitched together and also a good number of duplicates that we had acquired from different sources. When I sorted this all out, I had 3,346 books, and I ended up having to completely rewrite those last two chapters.



Sadly, one of the really cool and sexy things I thought I had discovered (on the left) turned out to be an aberration of that bad data--it completely disappeared along with my five-page analysis of why it was there in the first place. But honestly, I was not disappointed. This is how progress goes; we need to be open to the possibilities of error and of failure. Mistakes will be made (notice my optimistic use of the passive tense there). My hope is that those mistakes will be revelatory.

#### FLANDERS:

This is really interesting: in this case, you as the proprietor of the data were in a position to discover the error. It seems to me that there's a very interesting epistemological problem, especially for researchers working with collections they don't own: when your data shows you something you didn't expect, what's the status of your surprise? (Is it a form of skepticism? Does it make you go back and check your data? Do you treat it as a motivating revelation and

move on?) There's a very interesting article by Paul Fortier in which he dramatizes a moment like this, where he saw a pattern that seemed wrong to him, and when he checked it, he determined that his analysis algorithm was wrong.

The computer system was able to produce in minutes frequency and distribution profiles for each of the important themes in the novel. **But the theme profiles did not lead to an acceptable interpretation of the text.** After some reflection I discovered why.

—Paul A. Fortier, “Twentieth-Century French Prose Fiction”, 1989

25

But it seems to me that there's a potential circularity or undecidability here, particularly in cases where both the algorithm and the data are sufficiently complex, or sufficiently hidden from us, that we're not in a position to be skeptical of them. In a universe where both our own interpretive position and also the accuracy or truth-value of the tools are open to question, do we gauge the helpfulness of our tools against our (presumed true) beliefs? Do we gauge the truth of our beliefs against the (presumed accurate) tool? What's the status of the unexpected in these cases? Maybe that's a rhetorical question at this point.

In which the word “interpretation” is finally uttered openly, and the pedigree of big data mooted in return. . .

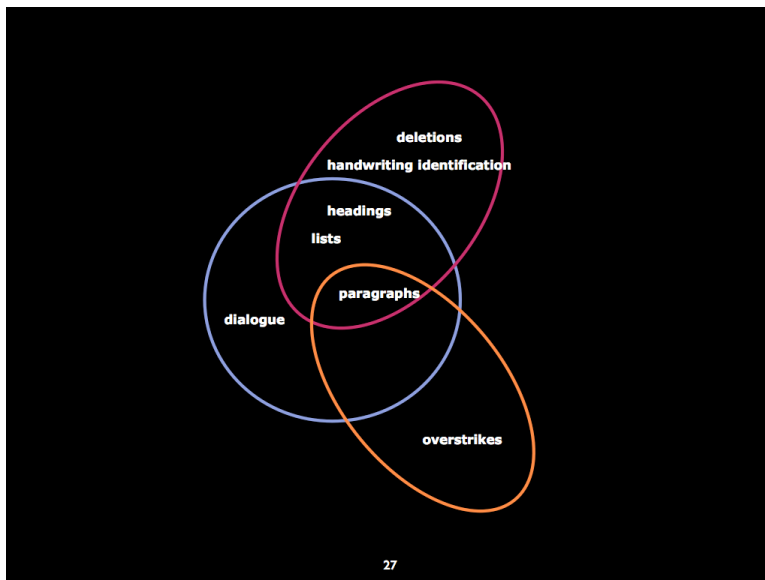
26

JOCKERS:

That question really leads me to the other thing that we really need to be mindful of: confirmation bias. This problem exists for us not only at the level of our interpretation of the data but at the level of our initial encoding and analysis of the data. As a programmer, I try to be as objective as I can, but I still have to schedule some time to think seriously about the extent to which I'm designing algorithms that are predestined to identify the kinds of things I want them to identify. I don't think I am gaming the system, but it's a possibility that I need to keep an open mind about. I wonder, Julia, to what extent you think about this problem and how we write ourselves into our own encoding practices?

FLANDERS:

There's no doubt in anyone's mind, I think, that markup constitutes a non-objective intervention in the text. In cases where it looks as if there's complete consensus about a markup decision, it's because we've chosen the boundaries of our consensus community in a certain way. But those forms of consensus are very powerful: I think as long as we understand their boundaries, it can be useful to treat them as having some truth-value, in practical terms.



So the question to ask about markup, I think, is “am I a member of the consensus community responsible for this markup?”—that’s the question that a user of markup would ask— and “what’s the consensus community for which I’m creating this markup?” (which is the question that a person creating markup should ask). A given document might well contain a base level of encoding intended for a very broad community, and then other forms of markup intended for more specific, limited communities, and there’s no reason why they shouldn’t coexist.

The one caveat here is that you need to be transparent about the meaning of your markup. There’s nothing wrong with using markup to make private, individualized observations about a text. What’s wrong is representing such observations so that they might be mistaken for something more broad-based. Sort of like Humpty Dumpty...

"There's glory for you!"

'I don't know what you mean by "glory",' Alice said.

Humpty Dumpty smiled contemptuously. 'Of course you don't — till I tell you. I meant "there's a nice knock-down argument for you!"'

'But "glory" doesn't mean "a nice knock-down argument",' Alice objected.

—Lewis Carroll, *Through the Looking Glass*

28

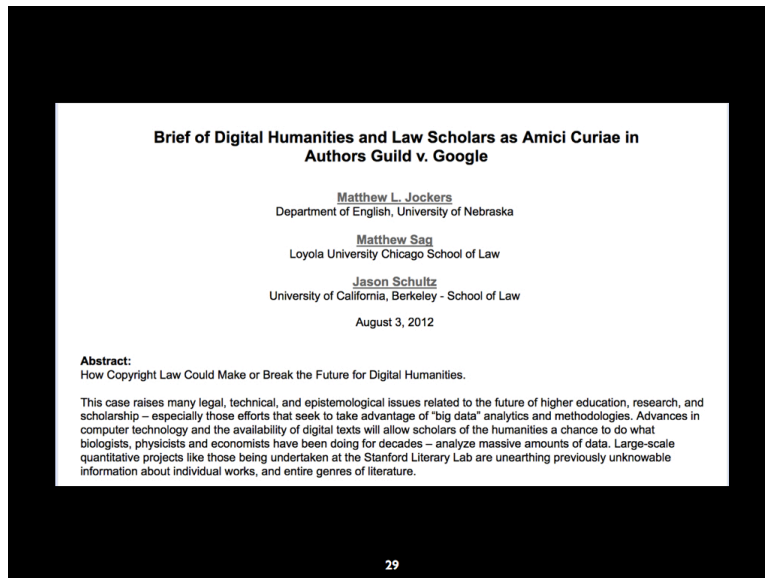
FLANDERS:

I also think there's an important social question that we haven't really addressed here, namely the ways that big data is funded. The creation of a very large data set (of any kind) is always going to represent a large investment, and in the current state of play, large investments are typically only made by organizations that have found a way to get a return on that investment. I have been really interested in the work that people like Marin Dacos are doing to establish the idea that data is a public good; I think that the climate in Europe may be more hospitable to this idea, though it's clearly an uphill battle no matter where you are. I wonder what the potential is in the future for large-scale data (e.g. on the scale of Google Books) that is developed as part of the academic patrimony, so to speak. You know more about this than I do—are there initiatives that are making progress in this direction?

JOCKERS:

Well, Julia, this is a pretty big can of worms. And, as you know, it's a can I opened several months ago by co-authoring a legal brief with Matthew Sag and Jason Shultz.





In the brief, we argue that the law needs to recognize the importance of “non-expressive” use of digitized content. The brief was admitted into the Authors Guild vs Hathi Trust law suit and judge Baer later cited it several times in his very favorable decision. Despite what I thought was a crushing victory for fair use and research, the Author’s Guild soon appealed the decision and so we are back in court.

Honestly, I don’t have the temperament or patience for this legal wrangling. To me the case is black and white, and if I say much more I’m likely to slip into the vernacular, as it were. I do think, however, that we must continue to work with what we have right now and not get hung up on the “if we don’t have it all we should not do anything” line of thinking. In fact, one way that we can continue to apply pressure in the legal realm is by showing time and time again that this work we are doing is truly non-expressive, transformative use. In the brief, we spend a good deal of time pointing to examples of this kind of work. So my advice is to keep pushing the research forward even while you are keenly aware of shortcomings. When the lawsuit is settled, you’ll be ready to rerun the program, and if your initial observations and interpretations change, so be it. You’ll be in good company with folks like Ptolemy.

In which they dream of greater things  
to come. . .

30

#### FLANDERS:

So while it is not a perfect world (or a geocentric one;-), it seems to me that what we've got here is actually a very interesting research agenda, and also some questions we could kick out to all those smart people who are eavesdropping on us. So how about we wrap this up with a bit of daydreaming. . . .For instance, here are two things I'd love to see someone tackle in the next five years:

#### Dreams

- Study markup as a discursive system.
- Develop an XML publishing system that moves between micro and macro.
- Fix Copyright and Fair Use.
- Deal with the metadata problem in large archives.

31

- I want someone to start studying markup as a discursive system, including the ways that it expresses scholarly opinion, perspective, call it what you like. The TAPAS project is starting to amass a very diverse collection of TEI-encoded scholarship, and I want someone to start treating that diversity as information rather than as noise.

- I'd really like to see someone to develop an XML publishing system with an interface for interacting with text collections that moves gracefully between micro and macro without losing sight of either one.

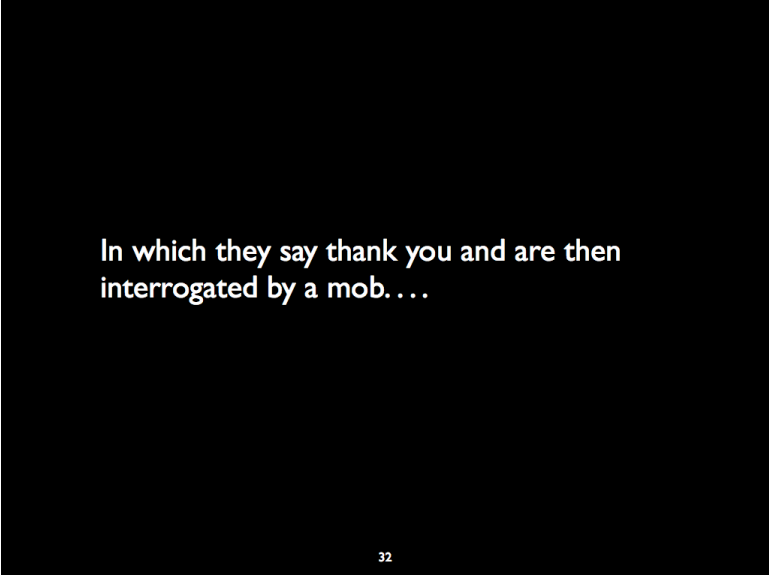
JOCKERS:

In terms of dreaming of the future, it's pretty hard for me to not to go right back to the copyright issue.

- Fix copyright and fair use!

but instead of beating that dead horse, I think

- I'd like to see us take a real serious stab at the metadata problem that exists within our large archives. The HathiTrust library is an incredible resource, but as it stands we can't even separate the fiction from the non-fiction. We can spend all the time we want building an text analysis platform, but until we know in a systematic and detailed way which text is which, I don't think the text-mining will be all that fruitful.



In which they say thank you and are then  
interrogated by a mob. . . .

32

JOCKERS: So, Julia, those are a couple of dreams. But let me conclude by saying what a real pleasure it has been dreaming with you here this afternoon and dreaming up this entire dialog. Thank you.

FLANDERS:

Yes, absolutely—thank **you**, and thanks as well to our indulgent audience. I think at this point we'd like to open up the dialogue!