Robert R. Stenson
Senior Thesis
Stephen Murray
04-05-2010

Churches Hidden in Plain Text

Mining and Mapping the Historiography of Gothic Architecture

Topic: developing a computational method for recognizing the names of Gothic churches as they appear in plaintext; using that recognition ability to search books for references to churches and so mine data from those books that can be represented on maps; finally, considering how such maps might influence the way we narrate the story of Gothic architecture.

This paper is the logical extension of a simple daydream: what if we taught a computer to think like a medieval cartographer? It is not such an odd question. Cartographers of the middle ages, unlike their post-Renaissance counterparts, did not make maps of the world as it appears from above.[1] Instead, they made maps from books (like the Bible), and rather than carving their maps into shapes and *spaces*, they filled them with ideas and points and *places*: names of cities, mountain ranges, names of ethnic groups—entities that, with the infinite patience of a monastic life, these cartographers had mined from text.[2] Maps were thus a way of "seeing" accumulated knowledge boiled down from books. Their maps were—and still are—answers to questions like: *what did the world of the Bible look like?*

But infinite patience is also a trait of modern computers, and it is not so difficult—with a

[1] Maps like those developed as navigational aids to a sea-faring Europe, and medieval Europeans were land-loving people (Delano-Smith 2006, 61). For a discussion of maps, see Harvey 1987, 465.
[2] For the importance of place (rather than space) in the medieval world, see Padrón 2004, 58. For a discussion of the text-mining behavior and general nature of medieval maps such as these (essentially limited to *mappaemundi* given the above description), see Harley and Woodward 1987, 326. Mary Carruthers also provides an example in her *Book of Memory*: Hugh of St Victor provides a world map in order to present his commentary *breviter* ("in brief") (294).

programming language and an internet connection—to turn a laptop into a *scriptorium* full of tabulating mendicants. A script searches a book for any reference to anything of interest, storing what it finds in a database. With that database we can create maps that let us "see" the book. This is the daydream: an updated version of the medieval cartographic-computational method for use on today's computers. And instead of using the medieval method to answer a Biblical question as above, this paper will attempt to answer a modern question: *what does the historiography of Gothic architecture look like?*

Why choose the historiography of Gothic architecture? Simply stated, the historiography of Gothic lends itself to the kind of text-mining described above. Its great histories are full of movement and references to churches; and Gothic churches themselves—composed piece-meal over many decades, often in divergent styles—seem to resist any one-way storytelling itineraries. When scholars tell the story of Gothic—in a textbook or a lecture or a slideshow—they lead their audience around France. Perhaps not in any real way (no buses or trains or walking tours), but still there is a common cognitive structure: *visiting*. We visit churches, discrete places on a map, and from these visits there slowly emerges a sense of relationships between churches. It is an integral part of understanding a phenomenon as diverse as Gothic architecture: the scholar must always be doubling-back, traveling the same roads twice, revisiting the Ile-de-France or some important cathedral, because coming to terms with only *one* Gothic church or buttress or façade always means coming terms with *many* examples of the same. Yet, when the scholar-storyteller sets down their narrative on a static map, as they often do, all of these well-developed cognitive relationships are lost. Each church becomes a simple point with two coordinates, and the stories told by different scholars of Gothic—even ones who disagree—begin to look the same. The journey evaporates, leaving out the rich and tangled trails, the reference patterns of the scholar.

Can a map in the medieval style rescue these traces and trails? (Hopefully the answer is yes.)

The second, more practical reason for addressing Gothic architecture is that this project is a small part of the much larger Mapping Gothic France project—from which I have drawn much for the code and database of this project, and to which I hope this project can contribute much in the way of code and data.

In order to answer the one big question—*what does the historiography look like?*—the paper itself is broken into three sections, each of which answers one of the questions: how, what, and why. (1) How do you build a system that can accurately recognize the names of churches as they appear in the wild, hidden like barely recognizable gems in historians' prose? This section is fairly technical, but both it and the project's code are written to make sense intuitively to non-technical readers. (2) To repeat the main question, *what* does the historiography look like when displayed on a map or a chart? This section covers a selection of animated maps and figures produced for the project from data generated in the first part. (3) *Why* does this "seeing" matter to the study of Gothic architecture? The final section is a fairly non-technical look at how understanding Gothic architecture and its historiography in terms of maps and networks and statistics might lead us to different understandings of the way we remember Gothic and put it together in narrative structures.

Part 1: Semi-Technical Description of Data Mining Scripts

Because this paper is meant for a humanities- rather than science-inclined reader, this section of paper—its most technical—will be short and to the point. In fact, the code written for this project follows the same logic: concise, and not overly magical/technical. What I have built is not a big black box that whirrs silently as it decomposes words into numbers and vectors (the kind of thing Google does when you are not looking). It is instead a relatively simple process that works with words and patterns. The magic of it is in not in its complexity but in the power of simple, labor-saving computations: programs that search for almost 400 churches in the form of 2000 aliases, all in a matter of seconds. The essential point: this project's algorithms for finding and learning names of churches, as well as this project's methods of manipulating that data, are designed to—at an abstract level—be understood by non-technical readers.

What is this central algorithm?[3] On the left-hand side of every Wikipedia page there is a link titled "What Links Here." If you click on the link, Wikipedia takes you to a list of articles that refer to the current article with a hyperlink. For instance, an article about Gargoyles links to the article about Amiens Cathedral, and the Gargoyle link refers to Amiens Cathedral as *Notre Dame d'Amiens*, rather than the more common *Amiens Cathedral*. To make sure that, when a user clicks *Notre Dame d'Amiens* in the Gargoyle article, the user will end up at the article about Amiens Cathedral, Wikipedia keeps a record of "redirects"—or "aliases" in the terminology of this paper and the Mapping Gothic France project—that make sure all traffic flows correctly. All of these redirects are viewable on the "What Links Here" page.

The first script in this project simply visits the "What Links Here" page for every church in a list and, after reading through each page, records all the redirects for that church. *Amiens*

[3] I hesitate to call it that, given it is a non-formal implementation of an idea that cannot be expressed easily in a bite-sized chunk of pseudo-code, the language of choice for algorithm designers.

Cathedral multiplies in an instant to be five more names:

Cathedral of Our Lady of Amiens

Notre Dame d'Amiens

Cathédrale Notre-Dame d'Amiens

Cathedral of Nôtre Dame, Amiens

Cathédrale d'Amiens

Each one of the aliases is considered equal to any other, no matter the spelling or seeming redundancy of names that differ only in diacritics. This project attempts to capture the *natural language* use of church names in all kinds of text, not the proscribed use or the one true way of referring to a church; *any* name is a good name. Next the script moves on to the French "Pages liées" and the German "Links auf diese Seite," searching the list of redirects and recording the variations. *La cathédrale d'Amiens* and *Kathedrale von Amiens*, for instance, are added to our list.

As you can see, Wikipedia is a kind of prism through which we can refract a single name and, on the other side, find the canonical white light broken into the spectrum of natural language variety: suddenly we can capture the astounding variety of names for a single building, and then, with a list of churches and a computer script, for ~300 more.[4] In a word, Wikipedia "multiplies" a limited amount of names into a much larger list.

Finding Church Names in Plain Text

After all this data is mined, the project database has "learned" quite a bit. Now it is ready to start recognizing the names of churches in plain text. For the recognition part of the project—slightly more complicated than the learning part—the strategy is to take all the aliases gathered

[4] The base list was culled from the Mapping Gothic France database. Famous German, Spanish, English, Swiss, and Portuguese cathedrals were added later in an attempt to add some international gravitas to the Francophilic project.

and count the frequency of individual words as they appear in those aliases. "Cathedral" appears 517 times in church name aliases, making it the most popular word used in referring to churches. On the other hand, "Amiens" appears only 22 times, and "Braine" appears only 11 times, making them the kind of words we are interested in, because they signal loudly and clearly that a specific church is being invoked in the neighborhood of that word. More specifically, if we find the word "cathedral" in a piece of text like this one—"[The] series can be described as follows: Fécamp east bays, Fécamp west bays, Rouen cathedral nave…"[5]—we will have to look through the 517 aliases mentioning "cathedral" before finding the one that fits *Rouen cathedral*. If we find the word "Rouen," however, we'll only have to try on a quick 19 aliases before finding one that fits.

Having counted the frequencies of these individual words in church name aliases, a script eliminates the most frequent ones, boiling the church names down to a list of "signal" words.[6] With this list of signals in memory, the script combs a text word-by-word, checking to see if the current word is on that list. If the script does come across a word like "Rouen," it tries to find the longest alias in the database that contains the word "Rouen" and occurs in the given text.[7] In this case, though the script tries "S-Maclou in Rouen" and "the Cathédrale Notre-Dame de Rouen," it determines that "Rouen cathedral" is the best—and only—fit.

Another example: consider the sentence, *Last summer the Mapping Gothic France team visited the abbey of St-Yved in Braine.* The script first spots *Yved*, and finds references to it:[8]

> *Église abbatiale Saint-**Yved** de Braine*
> *Église Saint- **Yved**- et- Notre- Dame*
> *Église Saint-**Yved** et Notre-Dame*
> *Église abbatiale de Saint-**Yved** de Braine*

[5] Bony 1983, 513.
[6] distillery.rb (see the code appendix)
[7] In computer science this is known as a "greedy" algorithm, because it wants to match the longest possible substring in the larger text. More technically, the code here operates by finding a signal word, then handing that word's specific collocation (it's surrounding word environment) to a separate function, which
[8] You can try this example for yourself if you visit http://gothic.b.uild.in/gs and you enter "St-Yved in Braine" as text into the text field.

*Saint-**Yved** (Braine)*
*Abbaye S-**Yved**, Braine*
*the Premonstratensian St **Yved**, Braine*
*St **Yved**, Braine*
*St **Yved** at Braine*
*St **Yved** Abbey in Braine*
*St **Yved***
*Église abbatiale Saint-**Yved** de Braine*
etc.

If you read the list, you'll notice that *St-Yved in Braine* is not actually there, yet the script correctly identifies it as equivalent to all the others. This is because the script is not matching to spelling or certain prepositions very strictly. Instead it tries to match the given text to a series of patterns[9] that allow variation in the exact spelling and wording of any alias. Because we have an alias for *St Yved at Braine*, a generalized pattern built from that alias allows, for example, dashes in place of spaces and many essentially equivalent words—*of*, *de*, *in*, *des*, *for*, *near*, or *aux*—in place of the single word *at*. In this way a simple set of rules can transform a name like *St-Yved at Braine* into a pattern that matches (and so recognizes) *St-Yved at Braine*-ishness (a whole family of names of a similar type), greatly improving this strategy for recognizing names in plain text.[10]

Conversely, a simpler way to recognize church names in plain text is to simply search for each alias in a given text, one-by-one, the way anyone searches for something on a web page or in a Word document. However, this method is much slower, as it wastes a considerable amount

[9] The patterns, known as Regular Expressions (given that they define Regular Languages, a kind of limited set of strings), look like this: `(s(|,|-|\.|\!|\?)|st(|,|-|\.|\!|\?)|st(|,|-|\.|\!|\?)|st\.(|,|-|\.|\!|\?)|st\.|saint(|,|-|\.|\!|\?)|saint(|,|-|\.|\!|\?)|sainte(|,|-|\.|\!|\?)|sainte(|,|-|\.|\!|\?)|ste\.(|,|-|\.|\!|\?)|ste(|,|-|\.|\!|\?)|ste\.(|,|-|\.|\!|\?))yved(.glise(|,|-|\.|\!|\?)|(|,|-|\.|\!|\?)(abbey|abbaye|church|abbatiale)?)?((|,|-|\.|\!|\?)|(|,|-|\.|\!|\?)|(|,|- |\.|\!|\?))?(of|de|at|in|des|for| near|en|aux)?((|,|-|\.|\!|\?)|(|,|-|\.|\!|\?)|d')braine` Though this may appear bewildering, it is actually quite simple. The | symbol is an "or" operator, so the pattern is really just saying "St. or St- or Saint or Ste.- or Ste. etc." and making some words, like "abbey" optional. For how these patterns are computed, see instillery.rb in the code appendix.

[10] This means my script accepts some fairly strange looking references to St-Yved at Braine, including st.yved.aux-braine and st.yved.ABBEY-in-Braine, just to name a few.

of effort in running through *every* alias in the database, rather than only a subset of the aliases.[11]

Internet Interface

An interactive internet interface to the database—available at http://gothic.b.uild.in—was put together to demonstrate the power of being able to immediately recognize the church names in plain text. Simply begin typing on the page, and the site silently combs what you type, looking for any church reference it can find. Once it finds church references, it plots those churches on a map and creates a tabulated list of each church mentioned and how many times the church has been mentioned. Though this may seem fairly prosaic, it seems to me there is real magic in being able to instantly turn simple written text—everyday prose without any annotations—into linked data like you might find on Wikipedia or any online encyclopedia.[12] Suddenly text can transcend its typesetting and reconfigure itself on maps and plans and charts and trees. Suddenly text can become data—structured, statistical, digital, reusable data—rather than just enjoyable knowledge that can be parsed only by scholars.[13] In another sense, such a program brings new life to the dreary process of hammering out a paper, because a semi-intelligent database is quietly reading

---

[11] Given the constraints of the searchable texts, this method was implemented, and is explained more thoroughly, in Part 2. Of course, the argument could be made that the simple searching is actually faster for larger portions of text, because—given my implementation of the signal word algorithm—the script could end up looking up the same aliases over and over again for a single piece of text. However, the signal word algorithm was developed basically to power the internet interface describe in the next section, which exposes short amounts of text to identification.

[12] Annotations-from-plaintext is a trope of the modern "semantic" internet, and is the driving idea behind such sites as evri.com or twine.com.

[13] To be more direct, this technology will be integrated with the Mapping Gothic France project as, in one use case, a way of building so-called "collections"—that is, of creating high-quality data sets that encompass discrete entities in the Mapping Gothic France system. Thus a user could start typing, "Jean Bony pits Bourges Cathedral and Chartres Cathedral against each other," and—without even trying—he will have a collection, ready-to-save, that includes both Bourges and Chartres. To my mind this is an elegant solution to the problem of a user building a collection, since this interface effectively eliminates the need for any interface at all save a text box—and what is more natural on a computer than typing text into a box? There has been work done on this area, specifically in building a system that turns written descriptions of scenes into actual 3D scenes. See Coyne and Sproat 2001 for a description of that system. In some ways, I hope, this seems like the future of computing: as computers become smarter and smarter with regards to natural language, we will most likely start to see all the twiddling knobs and controls of today's software start to disappear in favor of natural language interfaces intelligent enough to do all the knob-twiddling for us. The program available at gothic.b.uild.in attempts to be part of that natural language future.

and transforming text into pictures, drawing the scholar closer to the buildings, providing steady reminders of the visual forms on which he or she writes.[14] The compulsive writing synonymous with art history becomes, effortlessly, full of compulsive looking.

Other Notes on Name Recognition

Oddly enough, this pool of aliases is actually, by itself, an interesting body of knowledge, insofar as the number of aliases per church seems to be a (very) rough indicator of the church's hold in the popular—or even scholarly—imagination (fig.1). But can we take this as an indicator of the church's popularity or importance? Probably not, since the figure does not align that well with the data put together for Part 2 of this project, which tabulates the popularity of all churches across various books (fig.2).[15] There is consonance in the canon—Chartres has many aliases and is very popular in the historiography—but churches like S-Germain-des-Prés or Saint-Remi at Reims, which both have many aliases, are not too popular. Of course, the conclusion is inevitable and a little pedestrian: the names of these churches are difficult to spell, and so we end up with many failed attempts at correct spelling. It is interesting nonetheless to highlight the semi-genetic process at the heart of church names: the more a church is mentioned, the more likely it is that a name will "mutate." Even a church with a solid-sounding name like Notre Dame of Paris ends up with 50 invocations—*Notre Dame Cathedral of Paris*, *Notre-Dame-de-Paris*, *Notre-Dam in Paris*, *Our Lady of Paris*—simply because it is mentioned so often.

Also of note is that this final system for recognizing church names in plain text—the one

[14] This internet interface also allows users to add more aliases for churches if they notice that the program has missed a church name in the text. In this way the database of names is continually learning and getting better at name recognition.

[15] Although a logarithmic plot (as explored in Part 2) of the church names does seem to exhibit power-law behavior (i.e. it is linear on the log-log plot), the coefficient of its linearity seems to be entirely different from all other log-log plots developed, from word count in Wikipedia to church reference in Bony.

you can use interactively—does not at all adhere to the algorithm originally planned. Originally the program was designed to include a number of very modern technologies from the tool-belt of the modern computational linguist: automatic part-of-speech taggers, word sense disambiguation tools, etc. Early on, for instance, one version of this project included such a part-of-speech tagger in order to "chunk" texts—that is, look for general patterns like "Definite Article – Adjective – Noun – Preposition – Proper-Noun" (to match something like *the great cathedral at Amiens*). But while such a system does recognize church names with a high accuracy, it also matches any proper noun phrase with a high accuracy, leaving unsolved the problem: how can a computer differentiate between what is and what is *not* a reference to a church? (The algorithm outlined above solves this problem by being conservative: only finding aliases based on what it already knows. That is, while the above algorithm would match the phrase "The Church of Notre-Dame in Cleveland" and then attempt to match it to something in the database, this project's algorithm would simply ignore this reference altogether, never expending any energy trying to solve it.)

Still, there is nothing inherently wrong with these technologies; this project's accuracy might have improved greatly with their help. But there are problems. First, these systems are in general slower than what has been implemented here, due to both the computational complexity of their algorithms and the larger amounts of code they require to execute complex algorithms.[16] Second, these systems make assumptions and best-guesses with which most art historians would take issue.[17] Though 75% accuracy is cause for celebration among computer scientists, 25% inaccuracy is something for art historians to deride, and rightly so: art historians are not in the business of general pronouncements. But the final—and insurmountable—problem is that the

[16] It was the mantra of this project that "less is more" in terms of lines of code.

[17] As far as the state-of-art natural language technology is considered, > 95% accuracy on most artificial intelligence tasks is not in the cards for a few years. Of course, if such accuracy could be achieved, then the problem of art historians taking issue with the relative inaccuracy (~15% inaccurate) would disappear entirely. But my experience tells me that it is not in an art historian's vocabulary to forgive inaccuracies of any kind.

design of such complex tools is difficult to understand for non-technical and even technical readers alike (myself included). For a paper like this one, deferring the workload to black-box software toolkits seems to obfuscate the project's goal to, more than just accomplish its task, accomplish that task in an understandable way.

Difficulties in Name Recognition

To be honest, the task of recognizing church names in plain text was much more difficult than anticipated, and the accuracy of the system developed is low if considered among similar endeavors to "mine" information from texts. This project's code does not, for instance, capture any kind of indirect reference to a cathedral, as in "The architecture in Beauvais has this quality about it; the choir of the great cathedral and the nave of S-Lucien capture the style perfectly." My system would correctly tell you this text mentions Beauvais Cathedral (because "Beauvais" is an alias for Beauvais cathedral), but it would not capture the reference to S-Lucien, because the system is not intelligent enough to capture the earlier reference to Beauvais as a city and, with the knowledge that we are under the referential influence of Beauvais, infer that "S-Lucien" refers to S-Lucien of Beauvais. True, it would not be (very) hard to build a system smart enough to make such an inference, but incorporating such a guessing strategy would no doubt lead to an uptick in false positives, leading away from the conservative strategy implemented throughout the data-gathering process: quality over quantity when it comes to positive identifications.

Lastly, there is the problem of incompleteness: many cathedral and church names which we humans can recognize as church names—even if we know nothing about that church—are, for this system, impossible to recognize. The problem is that the system has no "knowledge" of those buildings, and, unlike a human, cannot recognize very general patterns. Jean Bony, in his

French Gothic Architecture, has a penchant for sprinkling obscure churches all throughout the text; even though most readers have not heard of those churches and though Bony provides no photographs, we still know they refer to actual churches. But because the primary seed for this project is the database of churches used in the MGF project, all those marginal churches remain unrecognized. Still, this deficiency is not such a loss. The churches not captured in the current program can, (1) be added later on, (2) do not have associated Wikipedia articles (breaking the prism mechanism noted earlier), and (3) live in the "long-tail" of Gothic architecture: buildings that occur very infrequently in the historiography and are of limited importance in the over-all network of inter-building influence. More often than not, these churches are dead nodes in the network: buildings that point to others, but to which no others point.

Part 2: Visualization of Gothic Historiography

The value of creating and maintaining a large database of names of Gothic churches may not be readily apparent; extremely long lists of nothing but names have the feel of data for data's sake, and the common contemporary practice of art historians—a solitary researcher absorbing current knowledge and, with intuition and analysis, producing a narrative argument—does not seem to make room for research based on such lists. True, careful analysis and cataloguing are still the common mode among archaeologists, as has long been the case, and it might helpful to think of this project as something similar: archaeology of the historiography rather than history itself. But the beauty of such a database as this is not simply that we can count the rows and columns and marvel at the sheer size and weight of all this aggregated data. The beauty of compiling these aliases—all the many ways of referring to many buildings—is that, when we search a text for Amiens Cathedral, we need not worry that "Amiens Cathedral" is an inadequate way of referring to it, because we are not simply searching for that *one* permutation of the building's name. We are searching for all 20 permutations of Amiens, ensuring that our search will find *all* references to Amiens qua Amiens, not Amiens dressed in one of many lexical guises. This tactic can be modeled as follows:

```
Given a text T,
For every alias A in the list of aliases of Amiens Cathedral,
        Search for A in T, counting occurrences of A
```

The result is a simple iterative search, first for *Amiens Cathedral*, then *Notre Dame of Amiens*, then *the cathedral of Our Lady of Amiens*, then *Notre-Dame d'Amiens*, etc. But if you've already asked a database to search the given text, why not search for more than simply Amiens? Why not search for Beauvais and Reims and Chartres as well, and the marginal churches too: Notre Dame of Voulton, S-Denis of Foulanges, etc. We need not bother asking an easy question (*Where do*

you mention Amiens Cathedral?) because the more interesting and complex question—*where do you make reference to* any *Gothic church, and how often to mention each one of them*—is just as easy as the first. To the loop-loving computer, the difference between running a search for 30 names and running a search for 3,000 is trivial, milliseconds apart in execution time. We need only expand our iterative search:

```
Given a text T,
For every church name N,
        For every alias A in the list of aliases of N,
                Search for A in T, counting occurrences R of A
```

To a human, however—the art historian hunched over a big book, tabulating references to each and every church—a question involving 200 referents and 3,000 references is laborious, tiring, and complex. Meanwhile, a computer can breezily expand this iterative search from a single book to a complete library; all we have to do is modify the first line of our algorithm to read not just "Given a single text **T**," but "For every text **T** in a library of texts **L**…" When faced with such a mountain of work, our imaginary (and ambitious) scholar is no doubt ready to heed that oft-given advice: perhaps you should concentrate on something more specific. But the idling computer is, as ever, happy to run whirr silently and compute without complaint, finishing one search and looping again, moving on to the next book on the shelf.

The net result is that having a large body of aliases allows us to *authoritatively* search a given text, to say that Jean Bony mentions Chartres qua Chartres 135 times in *French Gothic Architecture*[18] (which is an accurate count, according to my script) as opposed to his relative non-interest in Amiens, which only appears 77 times, despite its importance in Paul Frankl's[19] similar history of Gothic architecture, where it appears 126 times to Chartres' 121. Indeed, these

[18] Calculated from Bony 1983.
[19] Frankl 2000.

kind of comparative statistics form the core of what is possible with an authoritative search mechanism.

The problem underlying all text mining[20] is the problem of reference, explored in the last chapter, which was concerned with growing the database of aliases by "learning" new names of churches with generic patterns rather than direct matches. Here we are concerned with using that database of names against large amounts of text in order to calculate statistics. Search this, then that, then the other, then loop. (Rinse and repeat.) The result: we can begin to discuss general properties—the shape, the patterns, and the mathematics—of Gothic historiography taken as a whole.

Rather than concentrate on something art historically specific, this project aims to invite "computational thinking"—the application of automated, structured data analysis—into the art historical fray.[21] To be sure, this paper is *very* specific as far as computer science is concerned. To the computational linguist who trades regularly in the upper millions—millions of words in millions of documents—the domain of knowledge represented here is tiny: thousands of names of hundreds of churches scraped from tens of source texts.[22] But in the context of art history, this domain is impossibly large, non-specific to any one jamb statue on a façade, or a façade on a church, or even one church. The mode is *macro*, not *micro*—looking for patterns in the way we deal with many buildings all at once.

---

[20] The term means basically, capturing discrete "facts" in any text and so reducing a text, in some dimension, to tabulated data.

[21] For a discussion of "computational thinking" see Wing 2006, 33.

[22] Because of this specificity and small domain size, and because this is an art history paper meant to be accessible to art historians first and foremost, this paper does not deal explicitly with more of the mathematics of computational linguistics, despite ostensibly dealing with some very specific subtopics of computational linguistics, like named entity recognition, chunking, and power-law distributions in large corpuses. In general, the computational thinking employed herein owes less to the kind of structured experimentation common in university CS departments than it does to what software engineers commonly refer to as "hacking"—not in the sense of gaining illegal entry to a computer, but in the sense of making something on a computer quickly, primarily for the purpose of accomplishing something useful rather than creating some maintainable, efficient, or reusable (all of which are virtues in the world of software engineering).

Such a goal might seem a little grandiose, but the magic of computers is that an amateur art historian can attack long-standing art historical problems—albeit in a different dimension than a tenured professor—with freshness and abandon. As computer scientist Jeanette Wing explains: "Computational methods and models give us the courage to solve problems and design systems that no one of us would be capable of tackling alone."[23] Thus, in addition to developing a straightforward and sober system for finding church names in plaintext, this paper attempts to abandon the usual detail-oriented art historical studies and sets out, in the opposite direction, for the poorly explored country of data-driven research[24]—research that finds inspiration in creative manipulation and visualization of databases larger and more detailed than those any one historian could keep tidily in his head or papers. Even if, in the past, art historians have sunk their hands deep in data and tabulations, computational data is a different beast: quantifiable and visualizable in a way no *statistique monumentale* ever was. As art historian Maximilian Schich and physicist Albert Barabási have written: "In art history and archaeology… the increasing availability of massive amounts of quantitative data is fundamentally changing our perspective and research."[25] (This paper would like to take a small part in that change.)

But what exactly can a data-driven approach reveal about the network of Gothic churches that appeared in France from the 11th to 14th centuries? The general strategy for this project was to use the database of church aliases to authoritatively search a modest cross-section of Gothic

---

[23] Wing 2006, 33.

[24] This phrase "data-driven research" may seem overly opaque. The term is used predominantly in the sciences to describe a major paradigm shift brought on by the application of computational thinking to all areas of scientific endeavor, from physics to biology—a shift away from "hypothesis-driven research" (wherein a scientist tests some intuited hypothesis) toward "data-driven research" (wherein a researcher uses various computational methods to slice and dice data in an attempt to find patterns that might suggest, to that researcher, that there is some phenomena at work worth investigating further. For a general discussion of data-driven research, see *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Hey et. al 2009). To the best of my knowledge, one of the only scholars currently dealing with data-driven research in art history is the German Maximilian Schich, mentioned in the main text, whose work has dealt with power-law networks drawn from occurrences of figural motifs in classical art and scholarly references to various monuments in the Renaissance era. See Schich 2008 or Barabási and Schich 2009.

[25] Barabási and Schich 2009, 86.

historiography: digital books made searchable by Google Books[26] and articles made searchable by JSTOR.[27] Using data scraped from these resources, I developed visualizations of how the story of Gothic unfolds in and across these separate histories. Put it another way: this paper has produced many "images" of histories individually and together, and now we address the primary and beguiling question: *what does Gothic historiography look like?*

For example, what does Jean Bony's *French Gothic Architecture* (further considered in the next part) "look like"? In order to find out, I systematically searched a digital copy of the book made available through Google Books,[28] using a short computer script[29] that queries the Google Books data stream using the iterative algorithm discussed previously: every alias in the database, queried one by one. When the script asks Google Books to tell me where "Beauvais Cathedral" appears in the book, Google returns a list of snippets, among which is this one: "… Saint-Quentin and Beauvais Cathedral should be viewed as part of the Clairvaux derivation…" The script copies that and all the other snippets, saves them in a local database along with their page numbers and what church they reference, then moves on to the next alias in the database (which happens to be alias #1363, "Saint-Quentin").

With such data saved, any number of visualizations can be produced in order to image (or *imagine*) the book. For this project I produced three such images: an animated map, a linear plot of church-name references, and a "log-log" plot of the frequency with which the author mentions

---

[26] http://books.google.com. (Technically automated searching violates clause 2.15 of the Google Book Search Terms of Use: "you and your users will not…. use any robot, spider, site search/retrieval application, or other device to scrape, retrieve or index any portion of the Service (including, but not limited to, Google Book Search Content) or to collect information about users for any unauthorized purpose;" Because I was saving using an script to scrape information to a local database, this qualifies as "indexing," "scraping," and using a "robot." Fortunately, results from Murray 1989 were drawn from a local OCR'd copy of the text made available by the Visual Media Center at Columbia University, and as such do not fall under the Google Books part of my thesis results.)

[27] This part of the research used the very helpful Data For Research API made available by JSTOR at http://dfr.jstor.org/api. As far as I can tell, my data mining activity on JSTOR did not violate their Terms of Use.

[28] Bony 1983

[29] google.rb (appendix)

each building. Such visualizations were produced for all works studied.[30]

Animated Maps

The first visualization—the animated map—is certainly the most interesting from an intuitive and humanistic point of view, because it is the most familiar to art historians: churches distributed on a map. For these animations (available at http://thesis.uild.in), the viewer at first confronts a blank map of France. After the animation has started, a script[31] advances through a given book page-by-page, highlighting on the map any buildings mentioned on a given page, and drawing lines between that building and any other building that is mentioned on that page. The connecting lines are added to indicate that the buildings were mentioned on the same page, and while such a scheme may seem somewhat arbitrary, it was the most viable way of showing loose groupings in the churches (given the constraints of a Google Books search). When a building is mentioned more than once, the circle representing it on the map grows by 1 pixel, and—as the script marches onward page after page—oft-mentioned buildings bubble and flash like synapses or bomb-strikes. If Bony stops at a church frequently, its circle expands quickly, and it is these rhetorical *visits* that the animation captures effectively. We watch as Bony first casts a wide network of buildings across France—the important stylistic outposts of "Gothic"—and then, with that skeletal network in place, we watch him travel quickly from point to point, darting around France and filling out the network with marginal buildings, constantly re-visiting the buildings central to his story. (Bony's progress through the network will be explored in-depth in part 3). The final circles and lines left on the map—blipping no longer once the animation has stopped—

[30] These visualizations are available for every book or collection of articles studied on the project website: http://thesis.uild.in.)

[31] map.js—this code can be viewed "live" on the thesis website (thesis.uild.in) where it is being used to actually animate the maps currently.

are thus an "image" or impression of Bony's story of Gothic, taken all at once. His diachronic traipse comes together as a synchronic argument, pulled according to his preoccupations, and grown from data culled from his text (fig.3).

The use of such a map may not be readily apparent, but becomes so when juxtaposed with maps of other stories of Gothic, like Paul Frankl's *Gothic Architecture*.[32] Because Frankl (fig.4) is telling a story similar to Bony's, his network demonstrates a general similarity in shape, despite differences in local emphases: Amiens and Bourges cathedrals have reversed importance, and there is a general north-easterly attraction in Frankl's story toward Cologne Cathedral, at the expense of two westerly anchors in Bony's network, Chartres and Le Mans. (The pattern is not surprising, given that Frankl studied in Germany and Bony was born in Le Mans.[33]) There is, however, striking difference in how the networks develop. One can easily see, when watching the animation in slow motion, that Bony tends to lean on a certain church at any given moment in his story, stopping at that outpost and making excursions to surrounding churches: first at Saint-Denis outside Paris, then Laon Cathedral, then Bourges, then onto Chartres, a church that, once added to the narrative, blinks endlessly until book's end. What this animation shows is how Bony *remembers* the network as his tells his stories, because even as the network expands and the centers of concentration shift, the "true" center, Paris, is constantly a-light with references. When Bony moves on to a new building in his network, the building recalls another previously visited, and the recollection privileges buildings that have been visited often. Frankl remembers with a similar *habitus*, but he is more all-over—never settling down in one place, always keeping

[32] Frankl 2000.
[33] Idle speculation perhaps, but interesting nonetheless—and certainly the kind of observation that only emerges in raw data.

the entire network of Gothic structures at his finger tips.[34]

In the language of statistics, what we are seeing in both narratives is so-called "power-law" behavior, a kind of general trend in large-scale systems, and here taken to mean a positive feedback loop for prominent churches. Once a church figures into a narrative as an explanatory benchmark, the probability that it is mentioned again—that it is remembered and re-inscribed, told to blink in order to explain some *other* building—grows exponentially. In non-statistical terms: the rich get richer.[35] (This point is explored further in the section on plots.) Essentially, the magic of these animated maps is that we can describe with data the fascinating central axiom of architectural studies: that no building can be explained without the aide of others, because all buildings must exist in a network of similar ones, and that certain buildings become the centers of our attention, not because of any definable property of the individual *building*, but because of the way historical discourse is pre-stressed to privilege a canon.[36]

This point is born out especially in specialist studies of Gothic architecture, wherein the aim is to explain a part of Gothic rather than the whole. In Lindy Grant's *Architecture and Society in Normandy*[37], which deals with the region of Normandy (fig.5), the map skews heavily toward Rouen, Fécamp, and Coutances cathedrals, indicating without a doubt that Grant actually *is* emphasizing her ostensible subject matter. Yet the network we recognize from the previous

[34] In both instances, we can see that the narrative's movement is never constrained by space; every author darts back and forth freely. What should be explored further in this regard is whether or not the movement of the author is constrained in time. Ostensibly these are chronological histories of Gothic, but are they truly held in place by any chronology? This is a difficult question to answer, but given how disputed time is in relation to buildings, and how difficult it is generally to make a timeline of buildings, given how they develop in fits and starts, at times when different influences predominate. Part 3 will tackle this question.

[35] The "rich getting richer" parallel is from Barabási 2002, 81. Barabási is describing the phenomenon that 20% of the people in the United States control approximately 80% of the wealth.

[36] Maximilian Schich's own work on power-law behavior has focused on much the same idea in figural art of the Renaissance: the idea that a canon develops according to power-law behavior. See Schich, et al. 2008.

[37] Grant 2005.

maps remains. In Stephen Murray's *Beauvais Cathedral: Architecture of Transcendence*,[38] a monograph, the tremendous number of references (~1500) to Beauvais Cathedral swell enough to encompass all the other points on the map (fig.6),[39] but again the central network of Gothic—the conglomerations and emphases characteristic of Bony and Frankl—remains recognizable; the Gothic canon is insistent. In all of the maps studied, the environs of Paris blink brightly, the Ile-de-France becomes dark with reference, and a crescent of churches arranges itself south of Amiens: Rouen, Beauvais, Noyon, Cambrai, and Tournai. Observations like these may be old hat to a professor of Gothic, but to the amateur's, or even the non-specialists', eye, such a way of instantly "seeing" multiple narratives of Gothic might be a valuable shortcut to grasping large amounts of information quickly, a non-trivial advantage for scholars who utilize modern, information-saturated digital libraries. Surely it is easy to recognize a pointed arch or a flying buttress, or even to develop an eye for correlating style to period—certain tracery will always indicate Late Gothic, etc.—and begin to understand Gothic. But the stylistic is *one* dimension; complex history is another, and dealing with history's massive network of churches and dense historiography is a daunting task,[40] certainly a task never really stressed at an introductory level. Though we have notions of networks and networked maps in architectural history, particularly in Bony's work on France, the images of those networks are no doubt locked up in the minds of tenured professors or in the prose of books. That is to say: such maps are implied only, despite

[38] Murray 1989.

[39] Murray's network betrays the added curiosity of uncharacteristic interest—compared with the other maps—in Amiens and Troyes cathedrals. Of course, Prof. Murray has written books about both Amiens (Murray 1996) and Troyes (Murray 1987).

[40] Everyday our libraries are growing, journal articles are multiplying, and the prospect of ever becoming a specialist in more than the tiniest sliver of art history diminishes. Being able to conceptualize the full network of scholarship or a given style would be a tremendous antidote to information saturation, and harnessing computers to digest scholarship would mean being able to *embrace* information saturation rather than specialize endlessly as a way to, in my opinion, avoid the problem of information overload. It seems to me like a magical quirk in history that, at the same time as academia has grown beyond comprehension, new technologies are being developed to understand and harness the truly astounding explosion of knowledge.

their ostensible centrality to truly understanding Gothic in a historical dimension.

But so far we have talked about images of single works—the kind of image that perhaps, having read the book cover to cover, a careful reader will have imprinted on his or her mind. At one point in Bony's book he sets up a back and forth dialogue between Saint-Denis and Sens—something I noticed while reading the book, and I also noticed in the animation, which is to say: the animation does not *improve* our understanding of the book. But what if, rather than search just *one* book, we write a script to search—in an instant—50 years of scholarly articles from an art historical journal and so be able to create maps that advance year-by-year rather than page-by-page? What is the result?

Maps of scholarly journals exhibit the same patterns as individual books. In animated maps of *Gesta* (fig.7), *The Journal of the Society of Architectural Historians* (fig.8), and *The Art Bulletin* (fig.9), a theme develops: the story of Gothic as told by a group of scholars over fifty years exhibits the same general characteristics of the story of Gothic as told by one historian over a few hundred pages. The centers of interest move from place to place, but the canon—St. Denis and other churches living in the Ile-de-France—keeps attracting references from all around the network, year after year, page after page. On one hand is a coherent story of how *one* scholar understands Gothic, on the other is the story of how *many* scholars, at any given time, focus their attention in their attempts to understand Gothic, linking their stories to the greater network. Still, it is worth noting the obvious differences. In all the maps of these journals, there is both broader depth in marginal churches, and no real pattern to the diachronic movement—if anything, a map of a journal looks more like the final hundred pages of any one of the larger histories: the final pages in which an author darts quickly from church to church, in the throes of the full argument.

Plot of Churches-per-Page

The second graphic produced for this project is an attempt to capture how the narratives develop in time, without resorting to animation. Here church reference data is plotted against a given book's page numbers (the time dimension). The resulting plot measures the rate at which an author introduces buildings into the narrative (i.e., the number of new buildings per page). In other words, this plot shows how quickly and how deeply a historian moves through the network of churches. A steeper slope means a quicker movement through France, while a taller climb means more churches have been brought under consideration (fig.10).

Of particular interest are the clear differences between, again, the big histories of Bony and Frankl, and the specialist studies of Grant and Murray. For the big histories, there is a clear sweep of narrative that unites the entire book; new churches are constantly being introduced over the book's journey, and the languorous curve on the graph preserves the "tour" notion inherent in such wide-ranging studies—there is a single scholarly tour uniting the entire story. For Grant and Murray, however, there are clear breaks in their story structures. Both start with what appear to be quick introductions to wide swaths of the buildings under discussion, but then their rhetorical travels pause. (This has been represented as a return to 0 churches having been referenced, in order that the graph counts the same buildings as newly visited the next time a significant pattern develops.) Oddly, both Grant and Murray then make similar travels twice more—and at similar points in their books.

It is important to note that this single figure is only a distillation of more general images of patterns in each book. For each of the books a table was produced, in which a row indicates a church and a column indicates the page number (fig.11). A circle appears in a given cell if that

church appears on that page.[41] Such a diagram allows us to quickly identify where in a book a certain church gets its most attention, but also allows us—through the dragging and dropping of rows in the web interface—to identify larger patterns of movement from church to church (such as the previous figure on new churches per page).

In the case of journal articles, the similarity of general patterns is again striking, but the meaning is somewhat different here, as it was before: rather than page numbers, what we are seeing is year-by-year progress, essentially the measure of popularity of a church in any given year, but also a measure of its relevance to any study. Though no article may have been written on Chartres in 1972, there are still many references to it because it is a touchstone which many scholars are often touching on, if briefly, in order to explain something else.[42]

Log-Log Plots

Across all these graphics, however, one problem is insistent: they are mostly anecdotal: interesting and different in their perspective on age-old questions, but not rigorous enough to invite hypotheses that some *X* or *Y* is a definitive property of the way scholars narrate Gothic. They invite new ideas, but do not themselves offer any results.

One scholar who seeks to determine definitive properties of art historical phenomena is the researcher Maximilian Schich, whose work has explored data on the use of certain motifs in art, as well as references to monuments in scholarly works—a tact similar to this paper, but the emphasis in his work is reversed. Rather than measuring the number of ways in which one author references many monuments, his work is concerned with how many scholars reference only one

[41] In the web interface to these charts, each row can be dragged and dropped, allowing a user to interactively explore the diagram and uncover patterns such as those used in the previous figure.
[42] For considering year-by-year progress of journal articles, this "velocity" diagram is less useful, since in a journal there is no coherent story being told. Still, charts are available for these journal article collections online, and might very well have a use unseen by me.

monument. What Schich has found in various cases is that citations and references in historical networks adhere to classic laws governing other kinds of networks.[43] As Schich co-wrote with famed network researcher Albert Barabási: "[T]he reference patterns of art historians appear to follow the same hub dominated scale-free topology as the one characterizing the www, scientific citations, or the human cell."[44]

What does that mean in non-scientific terms? Essentially, they are referring to a concept known as power-laws, which describe many large group processes, such as word frequency in a corpus of text[45] or wealth distribution (the "rich getting richer" moment of earlier). The usage of words in the English language is distributed exponentially, meaning a few common words occur very often, while many normal words are used a respectable amount, and finally a "long tail" of uncommon words occur very infrequently. These classes of words parallel the upper, middle, and lower income brackets, each occupied by few, some, and very many respectively. It describes the process mentioned earlier as well: the probability that a scholar will visit a church, say Noyon, again in the narrative is proportional to the number of times Noyon has already been mentioned; though it may seem obvious, this means churches with a prominent place in the narrative will be used over and over again to explicate features of other churches when they are introduced into the story.[46] In network theory this is known as *preferential attachment*, in that the scholar is preferentially "attaching" his overall narrative to a handful of churches, which—as the larger circles on the maps indicate—are an author's signature: think of Chartres magnifying in Bony, Reims in Frankl, Fécamp in Grant, even Troyes and Amiens in Murray. (A computer can discern

[43] Schich et al. 2009, 5 and Schich et. al 2008, 4.
[44] Barabási and Schich 2009, 86.
[45] Though it is uncouth to reference Wikipedia articles in a humanities paper, and I have kept this reference out of the official bibliography, there is an excellent discussion of Zipf's Law, at wikipedia.org/wiki/Zipfs_law.
[46] This is not a fact, but is rather a hypothesis of this paper.

these signatures quite well.[47]) According to most network scientists, this kind of process lies at the root of most power-law patterns, or "scale-free" networks, helping us to understand how and why histories create new stories and still participate in a recognizable grand narrative uniting all histories.

In order to visually confirm the existence of such patterns, Barabási and Schich employ two diagrams. The first is a kind of spider-web network diagram (fig.12) difficult to understand except on an impressionistic level; a diagram that grabs you with an aesthetic of complexity, but does not reveal much in the way of intelligible results. The more helpful, though less seductive imagery associated with scale-free networks is the so-called *log-log* plot, whose x and y axes, rather than progressing linearly, progress exponentially. What on a standard linear plot appears as a classic exponential hockey-stick shape (global CO2 levels, human population, etc.), appears on a log-log plot as a more or less straight line, allowing researchers to better judge exponential processes, in that what often looks confusing and crowded on a linear plot becomes elegant on an exponential one. All of this a roundabout way of saying, once again: building references in these histories of Gothic seem to show signs of power-law behavior when plotted log-log (fig.13), and so the way we tell the story of Gothic starts to resemble all kinds of other processes.

Another common way of looking at power-laws like these is with the well-known 80/20 rule, also known as a Pareto distribution, which states simply—as regards wealth distribution—that 80% of the wealth in the United States, for example, belongs to roughly 20% of the United States population.[48] Similarly, we might find that 80% of the United States population lives in 20% of its cities, etc. For this project, that would means 80% of all references to churches refer

[47] I am referring here, vaguely, to an area of computer science knows as "Machine Learning"—drawing statistics out of texts and building algorithms that correctly associate certain statistics with other aspects of a text, like the author. The algorithms are refined to be quite good at recognition, and have helped answer questions about the authorship of the Federalist papers, for instance. See Jurafsky and Martin, 658.
[48] Barabási 2002, 67.

to only 20% of the churches mentioned in a given text. Does it hold?[49]

The table below tests the theory that 20% of the Gothic churches in France merit 80% of the attention of architectural historians. In the first column, the total number of churches is taken to be simply the number of churches referenced in the book, while in the second column, that total number of churches is estimated at 345, the total number in the MGF database. Thus the two statistics read as, in the case of Bony: 20% of the churches mentioned in his book claim 82% of his references in the first case, and in the second case, 20% of the total theoretical number of Gothic churches in France merit 81% of his attention.

	Total # Churches Mentioned in this book with margin of error	Estimated Total Number of Churches (estimated from Mapping Gothic France project)
Bony 1983	82%	81%
Frankl 2000	86%	84%
Grant 2005	77%	79%
Murray 1989	88% (77% when mentions of Beauvais are removed	90% (81% without Beauvais)
Gesta, 1963 – 2005	88%	82%
The Journal of the Society of Architectural Historians,1945 – 2005	80%	72%
The Art Bulletin, 1919 – 2005	88%	82%
Everything (Aggregate Count)	91%	94%

[49] A major problem with determining 80/20 rule values over this data is the basic question: how many churches are there in France that date from the Gothic period, and should that number be the background statistic—the 100% of churches—used in our calculation. For this calculation I did a simple assumption that the 100% churches in France was actually the total number of distinct churches referenced by the author, multiplied by 1.7 in order to account a probably large percentage of churches that my database of aliases did not and does not capture when searching a book. Of course, this is a guess on my part. Were we to keep a constant figure of between 200 and 250 churches, or even the full count of 345 churches represented in the database, as the 100% of churches that could *possibly* be mentioned, the figures change somewhat, but still hover around the classic 80/20 paradigm. The problem with all this calculating, however, is that I knew about the 80/20 paradigm before beginning the project, so it did not exactly reveal itself to me—I went looking for it and, perhaps wrongly, found it.

Given these results, I would say the 80/20 rule fits, if not precisely. Application of a rule like this is not a matter of dead-on precision, because numbers that hover around 80% clearly indicate that the general behavior of power-laws is applicable. And there is even knowledge to be gleaned from the margins of error. In Grant's case, the number is low, around 78%, and we know her book is slightly different than others in that it deals with a non-canonical part of the network of Gothic architecture. We might surmise, then, that in writing a book that emphasizes such an area, more churches are brought under heavy consideration (what the low 78% tells us), because these distinct Norman buildings must still be explained in terms of—connected to—the Ile-de-France High Gothic canon. In Murray's case, obviously, the extremely high value around 90% is an unequivocal indicator of the book's singular focus on Beauvais.

All considered, this paper is not so much about precise definition of power-law networks in Gothic historiography, but about the application of the alias-name database to *find* high quality data like that discussed here so far. Any number of descriptive statistics could be calculated and explicated from this data—the "distance" between buildings in terms of co-occurrence, the most commonly co-occurring buildings, the maximum degree of separation between two buildings, etc. [50] But the goal, all along, has been to demonstrate some potential directions in data-driven art historical research. More specific claims and calculations will be made in the next section, which focuses specifically on a computational critique of Bony 1983.

Quality of Search Data

A few problems in the data-collection process should be mentioned.

[1] Google Books limited search results for a given phrase to one-per-page of the targeted

[50] I will be the first to admit that my own general inability to do high-level mathematics slows down calculations such as these.

book; even if Bony mentioned "Chartres Cathedral" multiple times on page 313, for example, the search result would report that Bony mentioned Chartres *once* on that page. There is significant deviation therefore between the books I searched on Google and the one book I did not, Murray 1989, because my search of Murray would return all 13 matches on a single page, meaning the total reference count for Chartres—or, in Murray's case, Beauvais—is much higher than it would be had the book been searched on Google. This deviation most likely only effected the reference count for very popular churches, like Chartres and Saint-Denis, so the most popular buildings are not as popular as they should be.

[2] Nothing was done to ensure the quality of the reference found, meaning a reference to "Amiens" might actually be a reference to "the bishops of Amiens," which is not a true reference to the Gothic cathedral. There are now precautions in the system for rejecting matches that refer to the city rather than the cathedral of a city, but those precautions were not in place during the original searches of Google Books, a process now much hampered by Google's insistence on stopping my automated queries of their database. Also, in the case of JSTOR searches, often the articles returned were not actually about Gothic architecture, but rather simply mentioned some Gothic church in a non-Gothic context. Nevertheless, these deviations seem tolerable given that the false positives were not the bulk of the data, and that probably *all* the churches were simply scaled up by such a process, maintaining the basic character of the power-law pattern. Finally, given the correspondence in power-laws between general publications and more specific ones like the medieval-centric *Gesta*, it seems the noise was not too loud.

[3] Finally—and most importantly as regards network diagrams—is that there are two kinds of networks being described here: the network of references to specific churches, and the fairly un-researched network of inter-church relationships as claimed by authors: whenever an

author claims that church *x* is derived from church *y*, this would constitute a "directed" link in the network, an arrow connecting the influenced *x* to the influencer *y*. It had been the original goal of this project to actually capture and describe these kinds of relationships using technology from the field of natural language processing, and some progress was made in that direction, but the accuracy of the links as determined by computer was low—definitely too low in comparison to the rigorous precision of almost all art historical studies. Rather than attempting an inaccurate study, I determined it would be better to rely on co-occurrence as a loose, and basically informal, measure of "connectivity" in the network. All calculations rely instead on raw reference counts, which are not strictly speaking directed network data, but are still satisfactory given the overall behavior of the reference data on log-log plots, which is very similar to observations on word frequency in large text corpuses (demonstrated in the case of Zipf's law).

None of this is to say that a true directed network of influences would be useless—it is simply beyond the scope of this paper's natural language capabilities, and probably beyond the scope of *any* computational approach, given the attention to detail characteristic of art historians. Building such network visualizations would thus require heavy annotation of existing texts by trained human annotators, each one confident enough to draw discrete facts out of individual text, much like graduate students in linguistics who annotate text with parts-of-speech in order to further the work of computational linguists. Working with such data would be truly exciting, and in some sense the Mapping Gothic France project—of which this project is a subsidiary—will enable just that kind of high-quality data creation, but certainly not at the scale of any similar linguistics projects.

Part 3: Gothic in Reverse

So far this paper has considered many histories from a safe distance, skimming each in order to make observations about them all. But can this project's data illuminate a single work? Can an animated map help us investigate Jean Bony's *French Gothic Architecture*? In some sense this is an odd question: can an image of a book, ostensibly a faithful representation, tell us something about the book that the book itself cannot? But the question is valid, because the map of Bony's book contrasts so starkly with the book's language—its grand chronological sweep of Gothic couched in genealogical metaphors. It is language that conceals an architectural history whose predominant direction is not onward-and-upward but, by necessity, *backward*. What these Gothic structures preserve is not a teleology abutted by Romanesque and Renaissance, but the opposite: thousands of backward-looking memories captured in each and every feature of each and every building: a phenomenon network diagrams and computer memories represent with ease, but chronological histories struggle to capture. The map thus prods us to ask: if we can *visualize* Bony's history as a network, can we also borrow some ideas from the science of networks in order to critique his telling of the story of Gothic architecture?

In the early parts of his book, Bony peppers the narrative with handfuls of church names. The animated map records small sub-networks striking out into uncharted territory whenever he does this, and a reading confirms that these discursive lists are everywhere: four or five churches strung together, tracking the progress of a stylistic tendency as it makes its way, ghostlike, across France. S-Leu-d'Esserent, S-Germain-des-Pres, Chartres, and Bourges cathedral are invoked to recount early victories in the history of flying buttresses.[51] Le Mans, Troyes, Voulton, Provins, and Corbeil are said to descend from Lombard architecture.[52] A fuller example follows:

[51] Bony 1983, 42.
[52] Bony 1983, 71.

> From the Loire Valley it soon started spreading in a north and northwesterly direction toward Normandy and the Ile-de-France (choir of Bernay, Jumieges, nave of Mont-Saint-Michel in Normandy; nave of Saint-Germain-des-Prés in Paris), occasionally reaching as far as the Rhineland (nave of Speyer as redesigned ca. 1050)….[53]

Two features make these riffs stand out. First, the active verbs of stylistic spread always attach themselves to an "it"—a stylistic tendency, rather than any group of builders or a human agent—and this tendency always acts from past to present, moving forward in time and outward in space, like ink spilled on a map. Second, for many of the churches referenced, this is their first appearance in the book, and they are appearing as nothing more than a name-drop: no picture, no explication. What purpose do such names serve? For a professional scholar, they may contain real meaning—prods to recall in the mind the choirs and naves of the referenced churches[54]—but for an amateur there are no memories to call up. The names pass by meaninglessly. However, by the time Bony has reached page 200 and amassed a hefty network of buildings, such references take on meaning: at this point the churches referenced have already had their moment under the microscope, and now each simple name is a complex look *back* to what has come before, not just a bell rung for deaf ears. Still, it would be wrong to say Bony acknowledges this remembering. His active verbs persist even as the narrative begins to turn inward and backward. One church begets another, a formula conquers a region, time marches onward.

Something slightly different characterizes Bony's references to pre-Romanesque and pre-Gothic structures, however, because there is a pivot point in Bony's history: a blurry start point, before which buildings exist in a referential ether, available for citation rather than explication, and after which buildings exist as members of the stylistic genealogy currently on the examining table. The line in Bony seems to be around 1100, a date he attaches to a handful of north Italian

---

[53] Bony 1983, 83.

[54] The service available at http://gothic.b.uild.in does just that, although not for Speyer, which is outside the scope of the Mapping Gothic France photographic database.

Romanesque churches that exploit the style of vaulting essential to defining Gothic structures.[55] Accordingly, the Italian buildings merit brief examinations—around two or three lines each. But any building from considerably before the sentinel date of 1100 are nothing more than names that point ambiguously back to the shelves of architectural history books from which this one was plucked. For example, it goes without saying that, when he references "the large square bays of Roman groin-vaulted structures, such as the Basilica of Maxentius,"[56] Bony does not explicate that basilica any further. Doing so it would transgress the frame of his book,[57] and exploring such a basilica piece-by-piece might require continuing that exploration indefinitely: the roof of this basilica references a Greek building *y*, the Greek building *y* references some building *z*, etc. (It is the kind of scenario a resource like Wikipedia does not protect us from: you click on Basilica of Maxentius and, with only a few more clicks, you are reading about the Egyptian pyramids.[58])

The process described here is *recursion*, a notion popular in computer science and some linguistics that describes, to put it non-mathematically, a consistent deferral in order to answer a question or provide more information.[59] A common case in scholarship is a chain of footnotes. If you want to find out where Bony found the date 1099 for the church of Rivolta d'Adda, then you defer to his footnote. If the footnoted text footnotes another text, you defer once again, following the trail one step further, and so on, until one book in the chain makes a claim based on concrete evidence, at which point the recursive search ends and you can answer the original question—

---

[55] Bony 1983, 11.
[56] Bony 1983, 68.
[57] Where the north Italian churches do merit an accompanying photograph, the Basilica clearly does not. The assumption seems to be, (a) the reader is familiar already with this building, (b) can become familiar quickly, or (c)—most likely—if the reader is unfamiliar, it does not really matter.
[58] One might argue that Wikipedia has a refreshing lack of genre bias.
[59] In language the common example is a relative clause: *the façade, which was completed in 1284, stands…* The *which* clause is extra information that momentarily removes us from the main argument. In computer science the term is used most often to describe functions that are defined in terms of themselves. A certain function computes the factorial of a number *n* by invoking the same factorial-generating function on *n-1* and *n-2*, which recurs until a "base case" is reached, where the factorial of 1 or 0 is 1.

where did he get that date?—definitively.

Similarly, we find recursion in many historical digressions, like this section from Bony:

> Sens had only two little side chapels flanking the aisles: a peculiarity which was to be repeated soon after in the church of Saint-Père at Chartres but which was not unknown in the area, since it was present already in the choir of Saint-Thomas at Epernon… Actually the very type of these churches designed with an ambulatory but without transepts seems to have been a regional formula, particularly favored in the country south of Paris, between Seine and Loire, where it had begun a century before, in the Romanesque cathedrals of Chartres and Auxerre….[60]

Though it may not seem like the same process as footnote-chaining, Bony is constantly deferring, or *falling through* in this example. He begins with Sens Cathedral, takes a step forward to Saint-Père at Chartres, but then begins to look backward for examples of such naves, first to Saint-Thomas at Epernon, and then to a broader "regional formula": Romanesque cathedrals at Chartres and Auxerre. If we take Saint-Père as the starting point, we have a four-step recursion: Saint-Père to Sens, Sens to Epernon, and Epernon to the Romanesque cathedrals of Chartres and Auxerre. Afterwards Bony stops the backward motion, returns to Sens, and acknowledges his recursion: "[T]he cathedral of Sens should be linked to a whole background of still ill-identified tendencies."[61]

In reality, such recursion, or what we might call *discursive* narrative, is common.[62] An author, focused on one building, takes a step back in order to explain that building and push the main narrative forward. Then perhaps he takes another step back, and maybe another in order to fill-out some sub-argument. But he will always return to the main argument. For example, Bony structures his opening chapter as the tale of how rib vaults enter France. But because he begins

---

[60] Bony 1983, 65

[61] Bony 1983, 66.

[62] While the process is undoubtedly common, it seems almost non-existent as an explicit concept in historical writing. The only example of the concept in art history that I could find was a paper titled "Recursive Chaos in Defining Art Recursively" (Haines 2004), an impenetrably philosophical work combating the notion that art is art because it looks like previous works of art—a recursive notion apparently propagated by others in the field of the philosophy of art.

with north Italian churches, which are not the start of that style of vaulting, he is obliged to step backward in order to recount a chain of transmission that stretches all the way back to "distant oriental methods."[63] The chain terminates there.

Still, Bony's military and medical metaphors—victory, spread, etc.—far outnumber these rare recursive paragraphs. But what place do a political historian's metaphors have in a book so focused on stylistic evolution? The problem is that political and social history feature prominent *active* agents: armies, viruses, evangelizers of all stripes, each of them always pressing forward in time. Architectural history does have its own active agents, like the builders themselves, and from time to time Bony inserts them into his narrative: "Ever since the 1090s, the square Roman bay had haunted the imagination of the North Italian builders."[64] But there it is again! The ghost of Roman bays is a haunting *agent*; the Italian builders are merely acting on its behalf. And so, despite the builders' backward-longing, Bony has found a way to make them look forward in time.

The other active agent of stylistic history is, of course, a building. But we would be hard pressed to find pointers from a building to its "descendents." A building is built, it is finished, it is incapable of begetting. Discrete features of one derive from discrete features of another (not necessarily Gothic) building past, and it is in this derivation—this *remembering* of a particularly intriguing structural or decorative solution—that stylistic evolution takes place: invention occurs in a style's appropriation. Again, this is something the map captures. Bony enumerates features of one building, and other buildings light up around the map. When St. Denis is Bony's focus, St. Martin-des-Champs lights up twice and—if the map were built for it—Old St. Peter's and other

[63] Bony 1983, 13.
[64] Bony 1983, 68.

early Christian basilicas would light up too.[65]

But St. Denis is not only something tied to the past. Not only did the building continue to be built and to respond to the language of Gothic developing all around it, but it most certainly did "haunt" the minds of medieval Frenchmen; it is hard to shake off Bony's rhetoric. When we want to understand a building, we want to understand not only the building in stone—the sources of its styles—but also its effect on history. Similarly, when we read a book, we are interested in more than just the footnotes pointing backward, we might also like to peruse the (much harder to find) list of books this one has inspired. A building becomes important not because of anything inherent to the thing itself (save a formal feat like height or length or an exceptional detail, like Voulton's octopartite vault), but because its solutions and inventions can be heard echoing in the masonry of later structures or the spaces of the King's memory or manuscripts.

But who should keep track of all these links back and forth, these echoes and memories? Is it the business of St. Denis that the "scallop-like outline of the outer walls" of its chevet recalls the chevet of St. Martin-des-Champs,[66] or is it the business of St. Martin that St. Denis borrowed something from its particular decorative vocabulary? Put another way: which section of a history should link the two buildings—the section on St. Martin or the section on St. Denis? For Bony, it is not clear. His offhand lists of building references suggest that St. Martin should take care of charting its children. But surely St. Denis—the third most-mentioned building in all of Bony's text[67]—is too important a church to be merely a child of some less important building. So when Bony does focus on St. Denis, he does not hesitate to point backward to S-Martin-des-Champs. It is common throughout the entire book. Often, for every one step forward, Bony must take a few

[65] Currently the program only recognizes structures listed in the Mapping Gothic France database and associated with Wikipedia articles.
[66] Bony 1983, 68.
[67] This statistic is calculated from my project database, using information culled from Google's digital copy of the book. St. Denis is mentioned 107 times, Laon 115 times, and Chartres 135.

steps back, traversing the network once again in order to make sure everything connects. Indeed, what seems to separate canonical from non-canonical buildings—in one dimension—is the way Bony tells their story, either backward in time or forward in time. Great churches like St. Denis clearly merit a look back, while marginalia are glimpsed only in glances forward.[68]

For computers, however, only one of the two modes is efficient: it is the business of the later church to point back at its inspirateur, not the business of the inspirateur to keep track of all the buildings it has inspired. Representations of networks live in a computer's memory, and so are structured to make the most of that computer. Suppose we have a hypothetical network (like the one the MGF project is working to establish) that is interconnected with claims like "Bony says the chevet of St. Denis references the chevet of St. Martin-des-Champs". A good database representing the network would, without a doubt, prefer to associate this fact with St. Denis—"child" to "parent"—because the database can distribute pointers evenly among all the churches in a network this way. Any one church can point to only so many antecedents, because any one church has only so many discrete features to be anteceded directly. But if we take the opposite case—a "parent" pointing to a "child"—the responsibility of pointing weighs too heavily on the important churches; lists of children can grow infinitely. Consider the case of Chartres. Suppose every time a building points to Chartres we say the connection is from Chartres to child-x, as in this quotation from Bony:

> the propagation of the Chartres series can be followed, south to Orbais and Troyes, southeast to Reims, north to Saint-Quentin and Cambrai; spreading also soon after that from the Soissons area in a westerly direction to Amiens, Beauvais, and even to Rouen… or Les Andelys in eastern Normandy.[69]

But if the database makes its Chartres' responsibility to keep track of all its children, the

[68] It is this kind of linking that seems to give rise to the power-law behavior mentioned in part 2.
[69] Bony 1983, 255.

list of pointers for Chartres—one of the most popular churches in Gothic historiography—would outpace all the other pointer lists for smaller churches. Querying the database to find out what churches Chartres influenced would be easy (simply reading through the list of pointers), but finding out if a church points to Chartres would require plowing through Chartres' long list of "children" one-by-one, and just to confirm something that is a characteristic of the given child: that its plan has the flavor of the Chartrain scheme. Structuring the database around back-links, however, not only renders "parent"/"child" meaningless, it also captures the referential style of buildings themselves. Instead of reading off of a pre-determined list concocted by a scholar, the computer program must actually read the database records of every church in the network in order to determine the links, in much the same way a scholar must—with the image of Chartres available in memory—visit other churches to concoct such a list of "children."

But this is not only the most efficient way for a *computer* to remember and so represent Gothic. This is also the way *human* memory operates most efficiently, both for the mundane reason that it is easy to remember lots of short lists rather than a few long lists, and because these buildings demand we remember them in this specific, backwards way.[70] We visit each building in the mind and, looking around, allow each discomposed, discrete feature—the nave, the chevet, the "scallop-like outline of the outer walls"—to lead us backward in time to the contemplation of some other monument. When we reach that other monument, we discompose it in precisely the same way, piece-by-piece: consider the nave and the chevet and the "thinning out of all structural elements," to use more of Bony's language.[71] Each and every feature pulls us backward over the length and breadth of the Gothic genealogy, and as we move back in time, our search spreads out, recursion pulls us further, more buildings are visited, and we move away from the strictly

[70] Carruthers 2008, 217. She is making the point that remembering by *divisio* (remembering in small pieces) was a technique employed by medievals in order to maximize the capacity of their memory stores.
[71] Bony 1983, 62.

"Gothic"—now to Italy, to Muslim Spain, to Rome, etc. All of it comes from a simple visit. As Viollet-le-Duc writes in the opening of his description of Notre Dame at Dijon, "Let us now enter the church…"[72] This mental visiting was also a common trope in the medieval world of the mind. In her work on that subject, Mary Carruthers documents the role architecture played in the mechanics of remembering and recollection; in order to remember the sections of an oration or a book of psalms, buildings were used because they could be discomposed in discrete pieces, taken apart psalm-by-psalm.[73]

When we ask the question "Why does Rouen Cathedral look the way it does?," we have two ways of answering: (1) we start from the very beginning and work our way up to Rouen, or (2) we enter Rouen itself, we take a look around, and we break the one monolithic question into a handful of others: "Why does the frontispiece of Rouen look the way it does?" and "Why does the nave look the way it does?" and "Why do the base moldings look they way they do?" Each question pulls us backward, each answer invites more questions, and once we've answered one question, we can begin to tackle another. It is a process any programmer would recognize: a recursive way of defining something—defining a building in terms of its pieces and defining those piece in terms of other pieces of other buildings. In this way, Gothic buildings transport us backward in time—not only in the sense of stylistic references and not only by removing us from our present day, but also (in the case of Gothic churches especially) by immersing us in a style of building recognizable as radically different from our own.[74] When we task either a computer or our own minds with *remembering* the whole of Gothic architecture (as is the task of the Mapping Gothic France database and the memories of the scholars who have built it), we should allow the

[72] Viollet-le-Duc 1990, 93.
[73] Carruthers returns to this motif throughout her book *Craft of Thought* on, for example, page 35.
[74] As Murray writes, "Buildings—particularly medieval churches—have the power to transport the user *elsewhere* in time and space—whether retroactively to the monuments associated with Christ's life… or forward in time to the Second Coming…" (Murray 2009, 480).

algorithms of human memory to operate naturally: to fall back in time in order to understand and recite the stylistic history of Gothic architecture.

Simply open Bony at random—page 373—and come upon a church by chance, Troyes.[75] Bony highlights the way the clerestory absorbs the triforium, and notes that the choir of Saint-Remi at Reims earlier linked them with the use of continuous shafting. When we travel back to Saint-Remi on page 147, Bony notes its critical relationship with Laon, discussed at length in the previous pages, where he ties Laon to the nave at Tournai and the Romanesque chevet of Notre Dame of Soissons, a nave notable for its early adoption of pointed arches along the length of the arcade, a feature Bony connects, on p.19, with early pointed arches at Cluny III, a church first mentioned in connection with "Islamic architecture of the Near East."[76]

Understanding why Troyes looks the way it does means understanding why *all* of these buildings look the way they do. And not only that. It is also a backwards walk that avoids the parent-child determinism inherent in Bony's story. The problem with a parent-child metaphor is it imposes strict binaries (is-a-child/is-not-a-child) on Gothic and so denies this multifaceted remembering. Unlike organisms, one building does not have one or two distinct parents, it has *many*: each and every building that lived in the memory of each builder, of each patron, and of each clergy member. A structure can find genetic material anywhere, incorporating it at will, not as a function of some predecessors *will* to procreate. This multifaceted heritage argues strongly against the kind of Darwinian metaphors that vaguely inform Bony's work. But how else can we capture the story of Gothic? The evolutionary urge is persistent. It is easy to say that elevation types compete, that certain plan types survive. But the simplicity of the evolutionary metaphor—many branching from one—is simply inadequate, both computationally and rhetorically. So what

[75] Well not entirely by chance—I had to find a good example.
[76] Bony 1983, 17.

if we reverse the direction of these chronologies? Instead of writing a history of Gothic as an inevitable progression or conquering (which it is not), what if we write it as an inevitable falling backward (which it is), a consistent remembering?

The question being asked is, obliquely: can a narrative—a story of Gothic—incorporate or imitate the simple elegance of encoding Gothic in a computer's or a human's memory? If the buildings transport us backward in time, and *remembering* is an efficient linking device, why not *remember* the story? It is an odd, perhaps useless question. First of all, "remembering" histories are already common. Encyclopedias of Gothic, including the Mapping Gothic France project, are founded upon a recursive assumption: discrete articles are linked together without narrative glue, each blurb remembers others, and it is very possible to cut a path through an encyclopedia from a late Gothic structure to an early one. But writing a book like Bony's as a backward story would give us a tale robbed of all narrative art! And where exactly would such a story begin? Could a grand stylistic story begin with a study of late, derivative churches—dead-ends in the network of links? A database would say yes. But the excitement is in Gothic's conquest of height and width, as Bony might say, a chronological campaign. And yet, is the height of Beauvais only terrifying given that step-by-step conquest as outlined by Bony? Absolutely not! Experiencing the height of Beauvais is what makes it terrifying, and at what point in a narrative of Gothic is it absolutely necessary to introduce that terror of Beauvais' choir? When the perilous desire of later builders for even taller structures is inexplicable otherwise.

But here this meek and technical paper is exiting its comfort zone, and this essay-within-an-essay—originally meant only to demonstrate how network terminology might better explain Bony's narrative mechanics—will have to end with a final point. Whether or not a backwards-history is a ludicrous suggestion, we should listen for a central dissonance in the ways of telling

the story of Gothic: computer memories and human memories structure Gothic one way, while stories—like Bony's—structure Gothic an entirely different way. So it is worth asking: is there another way to keep track of the information locked up in narratives? A way of computationally *remembering* Bony's book that allows us, and future art historical researchers, to play with his notions in a way free of his rhetoric?

Consider a network graph like fig.14. It is a representation of a few formal relationships as outlined by Bony. But it could be extended, infinitely, to describe not only Bony's work, but also the work of Frankl and Grant and Murray. All you would need, in addition to the database of this project, is the phantom database mentioned twice already: a manually-created repository of links between the discrete formal features of buildings, a datastore within which the claims and the counter-claims of Gothic architectural historians could be formalized and understood with rigorous computational techniques. Concepts like stylistic groupings, "families" of churches, and even a "resistance" to Chartres (to steal one of Bony's phrases) might show up in visualizations of such data, and could be described and tested in a way no such formal claims—locked up in the art of prose—could ever be tested now. Most of all, however, such a database would emphasize that the historiography's true wealth of quantifiable information about Gothic architecture is, as of now, *unavailable* to modern scholars—not because the local library does not have a copy of every book a scholar might ever want to read, but because reading *every single book* in a given discipline (or even subdiscipline) has already become impossible. The critical contribution of such a database—or any way of encoding architectural historians' claims—would be to make that knowledge *available* to everyone, not just specialists in a sub-sub-discipline. Scholars could begin to deal with more information than they do today—more buildings, more time periods, more claims, more evidence; every iota at their finger trips, every idea indexed and searchable in

a matter of milliseconds, no matter how great the database swells.

For the entire history of architectural history, prose-built books have been the endpoint of any idea's lifecycle. Perhaps an idea is rescued and recycled or challenged from time to time, but from a book it came and to a book it shall return. All along, however, this paper's goal has been to transcend that lifecycle: to coax even the smallest bit of data out of prose and into the rows and columns of a database, where it can be reused and recycled and mapped easily. The goal all along has been to transform the great stories of Gothic into some kind of "between" state: a state in which data can be understood at a glance—on a map, on a chart—and yet retain the signature of its author. The ideal scenario: a professor presents the façade of Amiens and Bony's story of Gothic in much the same way, as images—the first as a picture, the second as a map, and both with similar caveats: though each is an imperfect representation of the thing it represents, each is still valuable because it captures and conveys the real excitement of a truly grand enterprise: a grandiose building and a grandiose book. After all, a Gothic church and a book about Gothic churches are not so different. Both attempt to construct, from pieces of buildings past, a cohesive story. Both Abbot Suger and Jean Bony wanted to synthesize a great amount of knowledge about architecture, from all parts of France and Italy, under one roof.[77] So if by now we have answered that central question (*what does the historiography of Gothic architecture look like?*) why not invite that question's answer—the creation of images to represent stories—into the near-darkness of the lecture hall, into the slideshows of buildings and statues?

To me this is an exciting thought, though it is only one more daydream.

[77] It is worth noting here Mary Carruthers' point that the medieval buildings were considered, in some way, to encode stories. The building was a representation of a the holy story, not simply a place to hear the holy story read aloud.

Appendix

The code below is presented so that, in the unlikely scenario that someone would like to read the code, they would have some hints as to how everything worked. Also, the code reproduced here is only about 1/5 of all the code produced for the project, but does represent the "smartest" parts of the code.

instillery.rb (code for turning a building name into a generic pattern)

```
require "rubygems"
require "sequel"
require "pp"

class Instillery

  @@transforms = [
    ["(-| |,)(of|de|at|in|des|for|near|en|aux)?(-| |d')",
                          "(-| |,)?(of|de|at|in|des|for|near|en|aux)?(-| |d')"],
                  ["(.glise | abbey)","(.glise | abbey)?"],
                  ["(n.tre(-| )dame( cathedral)?|our lady)",nil],
                  ["(the )?cat(h)?.dral(e)?","((the )?cat(h)?.dral(e)?)"],
                  ["(s-|st |st-|st\\. |st\\.|saint-|saint |sainte |sainte-|ste\\. |ste-|ste\\.-)",
                          "(s-|st |st-|st\\. |st\\.|saint-|saint |sainte |sainte-|ste\\. |ste-|ste\\.-)"],
                  ["(abbey|abbaye|church|abbatiale)","(abbey|abbaye|church|abbatiale)?"],
                  ["( |,|-)","( |,|-|\\.|\\!|\\?)"]]

  def initialize(db)
    @db = db # just a database reference
  end

  def instill(new_alias,canonical,saver)
    alias_record = {
      :name => new_alias,
      :canonical => canonical,
      :type => "building", # legacy?
      :source => saver
    }
    exists = false
    @db[:aliases].where(:name => new_alias).each { |r| exists = true }
    if exists
      return "already exists"
    else
      @db[:aliases].insert(alias_record)
      alias_id = nil
      # get the id back
      @db[:aliases].where(:name => new_alias).each { |r| alias_id = r[:id] }
      # need to save it and get it back
      pattern_record = {
        :pattern => self.transform(new_alias),
        :alias => alias_id,
        :canonical => canonical
      }
      @db[:ruby_patterns].insert(pattern_record)
      return "it has been saved"
    end
  end

  def transform(text)
    text.downcase!
    @@transforms.each do |tf|
      tf[1] = tf[0] if tf[1] == nil
      text.gsub!(Regexp.new(tf[0],Regexp::IGNORECASE),tf[1])
    end
    #text.gsub!(/\303\251|e|\303\250|\303\252|\303\211|o|\303\264/,
     # "(\303\251|e|\303\250|\303\252|\303\211|o|\303\264)")
```

```
    return text
  end

  def copy_over(text,a_id,c_id)
    record = {
      :pattern => self.transform(text),
      :alias => a_id, :canonical => c_id
    }
    @db[:ruby_patterns].insert(record)
  end

  def transform_all_aliases
    @db[:ruby_patterns].delete
    @db[:aliases].each do |a|
      self.copy_over(a[:name],a[:id],a[:canonical])
    end
  end

end

#code for transferring all aliases
#db = Sequel.sqlite("../database/thesis.db")
#instillery = Instillery.new(db)
#instillery.transform_all_aliases
```

distillery.rb (code for mining building names from plaintext)

```
require "rubygems"
require "strscan"
require "sequel"
require "pp"

class Distillery

  def initialize(db)
    @db = db
  end

  def distill
    words = {}
    @db[:aliases].each do |a|
      name = a[:name].downcase.gsub("\u2019","'")
      #name = name.gsub!(/\303\251|e|\303\250|\303\252|\303\211|o|\303\264/,".")
      name.split(/( |-|,|\.|'|\(|\)|!|\?|;)/).each do |w|
        if words.has_key?(w)
          words[w].push_pointer(a[:id]) # attach to GoodWord
        else
          words[w] = GoodWord.new(w,a[:id])
        end
      end
    end
    distills = []
    words.to_a.each do |w|
      if w[0] == "eu" # should be dame, just testing...
        distills.push(w[1])
      elsif w[1].count < 150 and w[0].length > 3 and w[0] != "cathedral"
        distills.push(w[1])
      elsif w[1].count < 10 and w[0].length == 3 and w[0] != "and"
        distills.push(w[1])
      end
    end
    return distills
  end

end

class GoodWord

  attr_accessor :word, :count

  def initialize(word,canonical)
    @word = word
    @pointers = [canonical]
    @count = 1 # how many times does the GoodWord appear?
    @has_been_read = false
  end
```

```
  def push_pointer(canonical)
    @count += 1
    @pointers.push(canonical)
  end

  def pointers
    if @has_been_read == false
      @pointers = @pointers.uniq
    else
      @has_been_read = true # make sure we don't unique it again
    end
    return @pointers
  end

end

class Interlocutor

  # @db is a database
  # @text is a string
  # @bock is an array of GoodWord objects

  @@breaker = Regexp.new(/( |-|,|\.|'|\(|\)|!|\?|;)/)

  def initialize(db, text, bock)
    @db, @text, @bock = db, text.gsub("\u2019","'"), self.good_hashify(bock)
    @finds = []
  end

  def good_hashify(a)
    Hash[ *a.collect { |k| [ k.word, k ] }.flatten ]
  end

  # just a function that latches onto important words and shows
  # aliases they might refer to

  def preview(text)
    lookups = []
    text.split(@@breaker).each do |unit|
      if @bock.has_key?(unit.downcase)
        self.aliases_like(unit).each { |a| lookups.push(a) }
      end
    end
    return lookups
  end

  def aliases_like(word)
    @db[:aliases].where(:name.like("%#{word}%")).collect { |r| r[:name] }
  end

  # actually replace stuff in the text with meaningful brackets

  def interlocute
    scanner = StringScanner.new(@text)
    matches = [] # things we found, replaced in text with numbers
    rejects = [] # bad stuff we found, to be re-placed at end
    start = 0
    while scanner.scan_until(@@breaker)
      word = @text[start..scanner.pos-2].gsub("\u2019","'").downcase
      #word.gsub!(/\303\251|e|\303\250|\303\252|\303\211|o|\303\264/,".")
      if @bock.has_key?(word)
        best_guess = self.guess(@bock[word],scanner.pos)
        if best_guess
          matches.push(best_guess)
          index = matches.length.to_s
          @text.sub!(best_guess[0],index+"_"*(best_guess[0].length-index.length))
        end
      end
      start = scanner.pos # move back pointer along
    end
    @text.scan(/(([0-9]{1,})_{1,})/).each do |m|
      match = matches[m[1].to_i-1]
      if self.accept?(m[0])
        @text = @text.sub(m[0],"((#{match[0]})[#{match[1]}])")
      else
        @text = @text.sub(m[0],match[0])
      end
    end
    return @text
  end
```

```
  # figure out the longest, most appropriate match

  def guess(good_word,position)
    backset = 55 # default
    if position < backset # check if default is too far back
      backset = position - 1
    end
    collocation = @text[position-backset..position+35]
    potentials = []
    good_word.pointers.each do |pointer|
      @db[:ruby_patterns].where(:alias => pointer).each do |r|
        pattern = r[:pattern]
        #pattern = self.fix_unicode_for(r[:pattern])
        match = collocation.match(Regexp.new(pattern,Regexp::IGNORECASE))
        potentials.push([match[0].strip,r[:canonical]]) if match != nil
        #puts pattern if match != nil
      end
    end
    potentials = potentials.uniq.sort { |a,b| b[0].length <=> a[0].length }
    return potentials[0]
  end

  def accept?(m)
    if m.length > 10 # lame assumption, but quick!
      return true
    elsif @text.match(/(the (bishop(s)?|town|city|area|region) of )#{m}/i)
      return false
    elsif @text.match(/(north|south|east|west) of #{m}/i)
      return false
    else
      return true
    end
  end
```

google.rb (code for searching Google Books automatically for all churches)

```
require "rubygems"
require "open-uri"
require "sequel"
require "json"
require "pp"

base = "http://books.google.com/books?"
@db = Sequel.sqlite("thesis.db") # the database!

def build_book_url(id,query)
  url = "http://books.google.com/books?"
  url += "id=#{id}&q=#{query}"
  url += "&pg=PA230&ei=jegSS6iVE6aQyAT4pMycDQ"
  url += "&jscmd=SearchWithinVolume&scoring=r#v=onepage&f=false"
  puts URI.encode(url)
  return URI.encode(url)
end

def book_query(id,query)
  url = build_book_url(id,query)
  result = open(url).read()
  json = JSON.parse(result)
  return json
end


def do_search(name,a_id,c_id)
  bony = "k7ytJ-gXonMC" # the unique book id for Bony's book considered here
  results = book_query(bony,"\"#{name}\"")
  if results["search_results"]
    results["search_results"].each do |result|
      if(result["page_number"].to_i < 515) # avoid the bibliography
        line = result["snippet_text"].gsub!(/<b>|<\/b>|[\d]+/,"")
        record = {
          :text => line,
          :alias => a_id,
          :canonical => c_id,
          :page_number => result["page_number"].to_i,
          :author => "bony"
        }
        pp record
```

```
        @db[:histories].insert(record)
      end
    end
  end
end

confirms = {}
#@db[:aliases].where(:canonical => 1164).each do |a|
@db[:aliases].where(:canonical => 9999..11000).each do |a|
  if confirms.has_key?(a[:name].downcase) == false
    confirms[a[:name].downcase] = true # make sure we don't check it again
    if a[:name].match(/^(the )?N.tre(-| )Dame$/) == nil
      do_search(a[:name],a[:id],a[:canonical])
    end
  end
end
```

Bibliography

Barabási, Albert-László. *Linked: the New Science of Networks*. Cambridge, Mass.: Perseus Pub., 2002. Print.

Barabási, Albert-Lászlo, and Maximilian Schich. "Human Activity from the Renaissance to the 21st Century." In *Cultures of Change: Social Atoms and Electronic Lives*, edited by Gennaro Ascione, Cinta Massip, Joseph Perello. Barcelona: Actar and Arts Santa Monica, 2009. Web.

Belting, Hans. *The End of the History of Art?* Chicago: University of Chicago Press, 1987.

Bony, Jean. *French Gothic Architecture of the 12th and 13th Centuries*. Berkeley: University of California Press, 1983. Print.

Carruthers, Mary. *The Craft of Thought: Meditation, Rhetoric, and the Making of Images, 400-1200*. Cambridge, U.K.; New York: Cambridge University Press, 2000.

———. *The Book of Memory: a Study of Memory in Medieval Culture*. Cambridge, U.K.; New York: Cambridge University Press, 2008.

Coyne, Bob and Richard Sproat. "WordsEye: an Automatic Text-to-Scene Conversion System," in *SIGGRAPH '01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001: 487-496. Web.

Delano-Smith, Catherine. "Milieus of Mobility: Itineraries, Route Maps, and Road Maps." In *Cartographies of Travel and Navigation*, ed. James R. Akerman. Chicago: University of Chicago Press, 2006.

Frankl, Paul. *Gothic Architecture*. New Haven, Conn.: Yale University Press, 2000.

Grant, Lindy. *Architecture and Society in Normandy, 1120 – 1270*. New Haven, Conn.: Yale University Press, 2005. Print.

Haines, Victor Yelverton. "Recursive Chaos in Defining Art Recursively." *British Journal of Aesthetics* 44:1 (2004): 73-83.

Harvey, P. D. A. "Local and Regional Cartography in Medieval Europe." In *The History of Cartography*, edited by J.B. Harley and David Woodward, 464-492. Chicago: University of Chicago Press, 1987.

Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall, 2009. Print.

Murray, Stephen. *Beauvais Cathedral: Architecture of Transcendence*. Princeton, N.J.: Princeton

University Press, 1989.

———. "Romanesque and Gothic Architecture," in *A Companion to the Medieval World*, edited by Carol Lansing and Edward D. English. Chichester, U.K.: Wiley-Blackwell, 2009.

Padrón, Ricardo. *The Spacious Word: Cartography, Literature, and Empire in Early Modern Spain*. Chicago: University of Chicago Press, 2004.

Schich, Maximilian, Sune Lehmann, and Juyong Park. "Dissecting the Canon: Visual Subject Co-Popularity Networks in Art Research." 5th European Conference on Complex Systems, Jerusalem (5 Sep. 2008). Web.

Schich, Maximilian, César Hidalgo, Sune Lehmann, and Juyong Park. "The Network of Subject Co-Popularity in Classical Archaeology." *Bolletino de Archaeologia On-line*. 2009. Web.

Viollet-le-Duc, Eugène-Emmanuel. *The Architectural Theory of Viollet-le-Duc: Readings and Commentary*, edited by M.F. Hearn. Cambridge, Mass: MIT Press, 1990.

Wing, Jeanette M. "Computational Thinking." In *Communications of the ACM*, 49:3 (2006): 33-35. Digital.

Selected Figures (all figures available at http://thesis.uild.in)

Fig.1: **Churches Plotted According to Number of Aliases**
(the author)

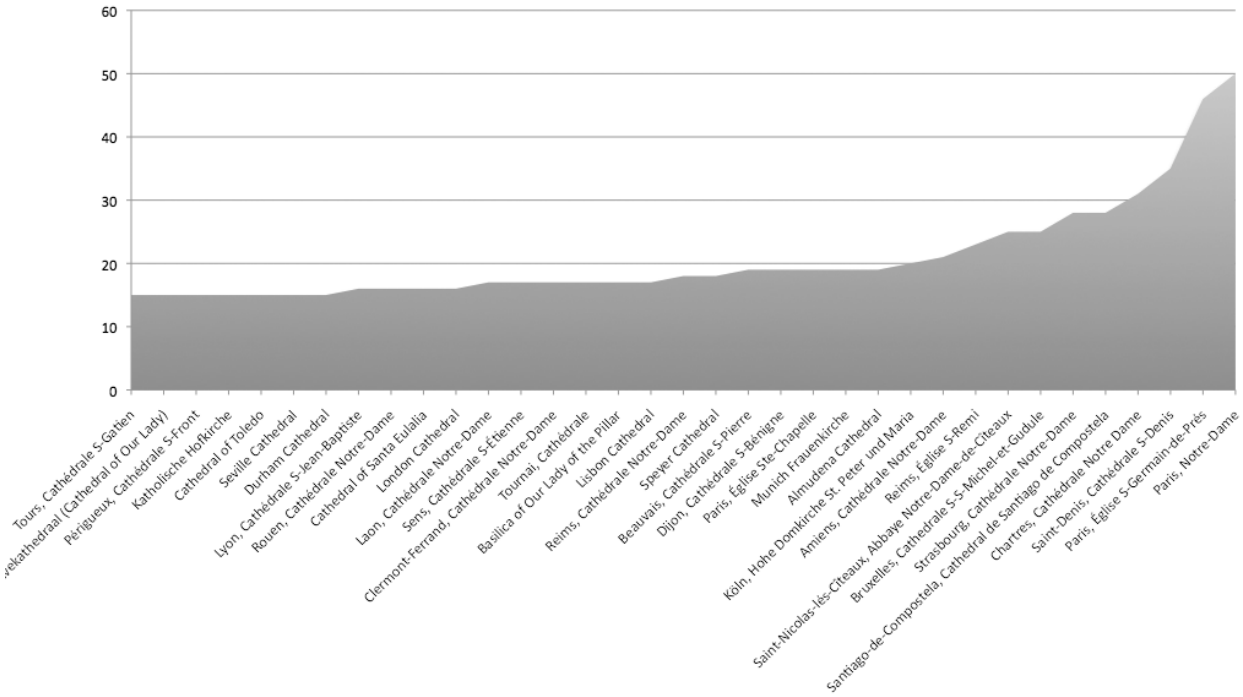

Fig.2: **Churches Plotted According to Number of References in Selected Historiography**
(the author)

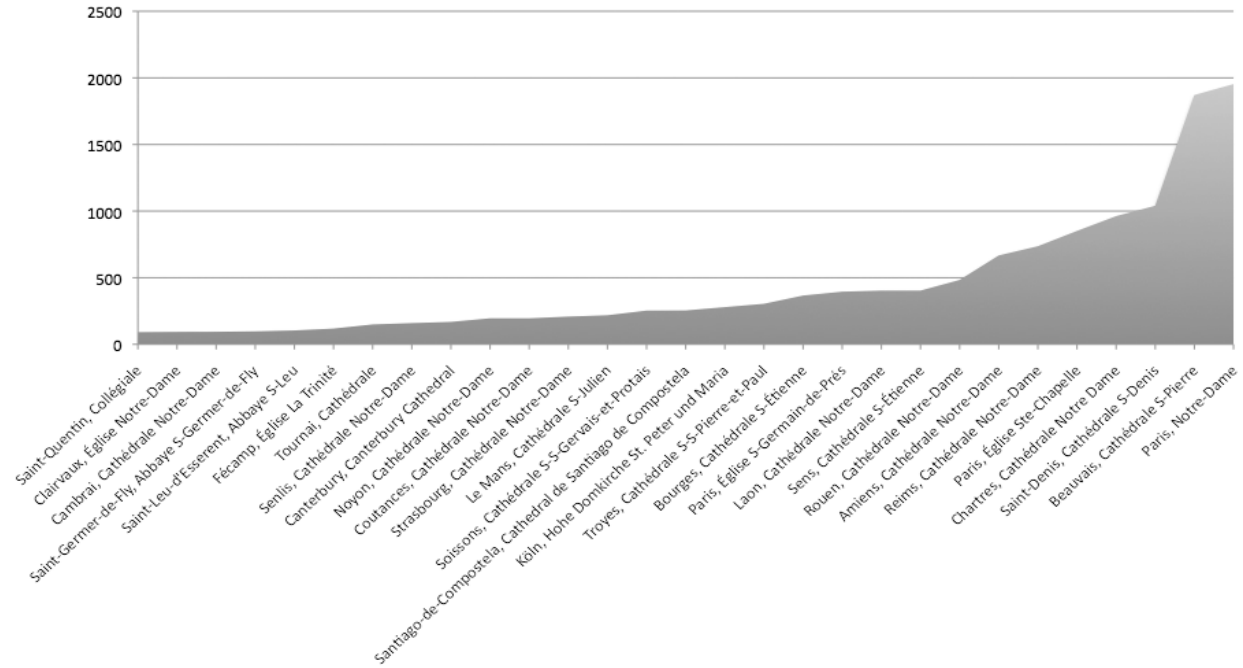

Fig.3: **Endframe of Animation of Bony's Map**
(the author; full animation available online)

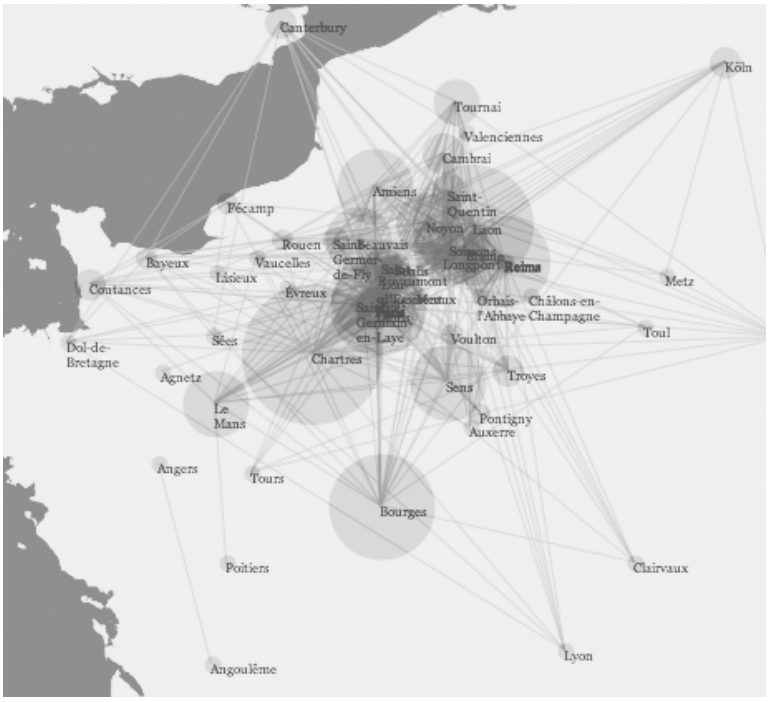

Fig.10: **Speed of Scholarly Movement Through the Network of Churches**
(the author)

This chart is explained fully in the text, but shows—basically—the speed at which an author introduces new buildings into the narrative. Frankl and Bony, writers of long histories, clearly unite their stories with long, slow introductions to many buildings, as indicated by the low slope of their lines compared to the early lines of Grant and Murray, who are writing specialist studies, not coffee-table sized histories.

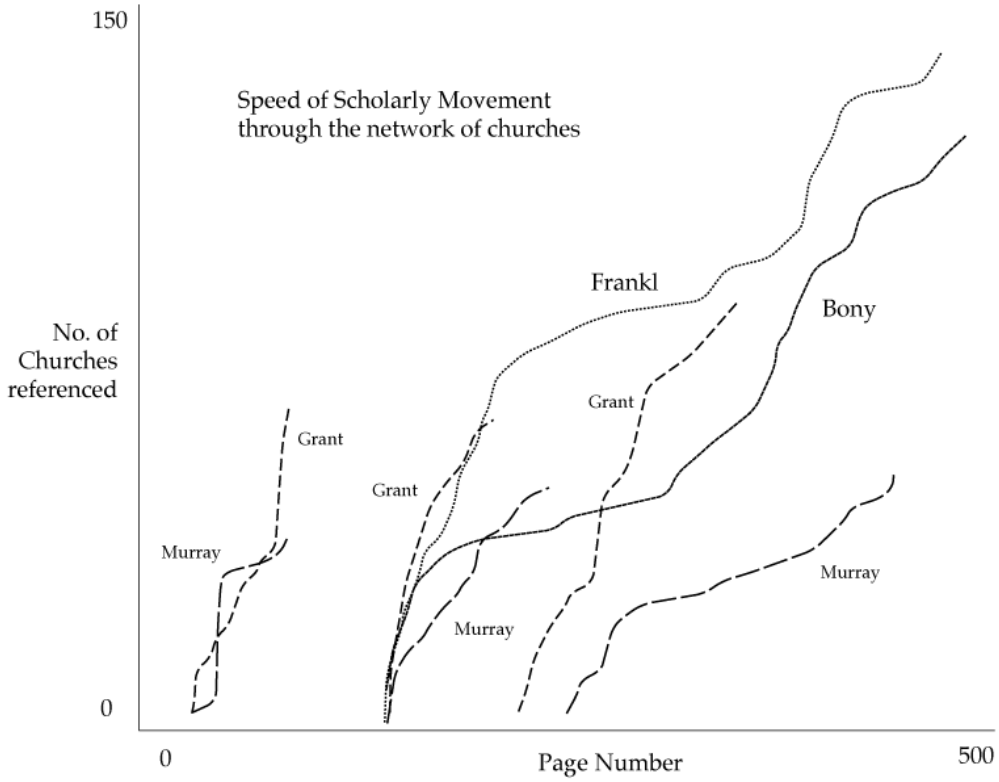

Fig.12: **"Reception of Ancient Monuments in Modern Scholarly Literature"**
(Barabási and Schich 2010, 89)

This illustration is included mainly as a example of the kind of illustration this paper was *not* aiming to produce. It is loud on visual effect, but quiet on meaning, especially without a key (and one is not provided by Barabási and Schich).

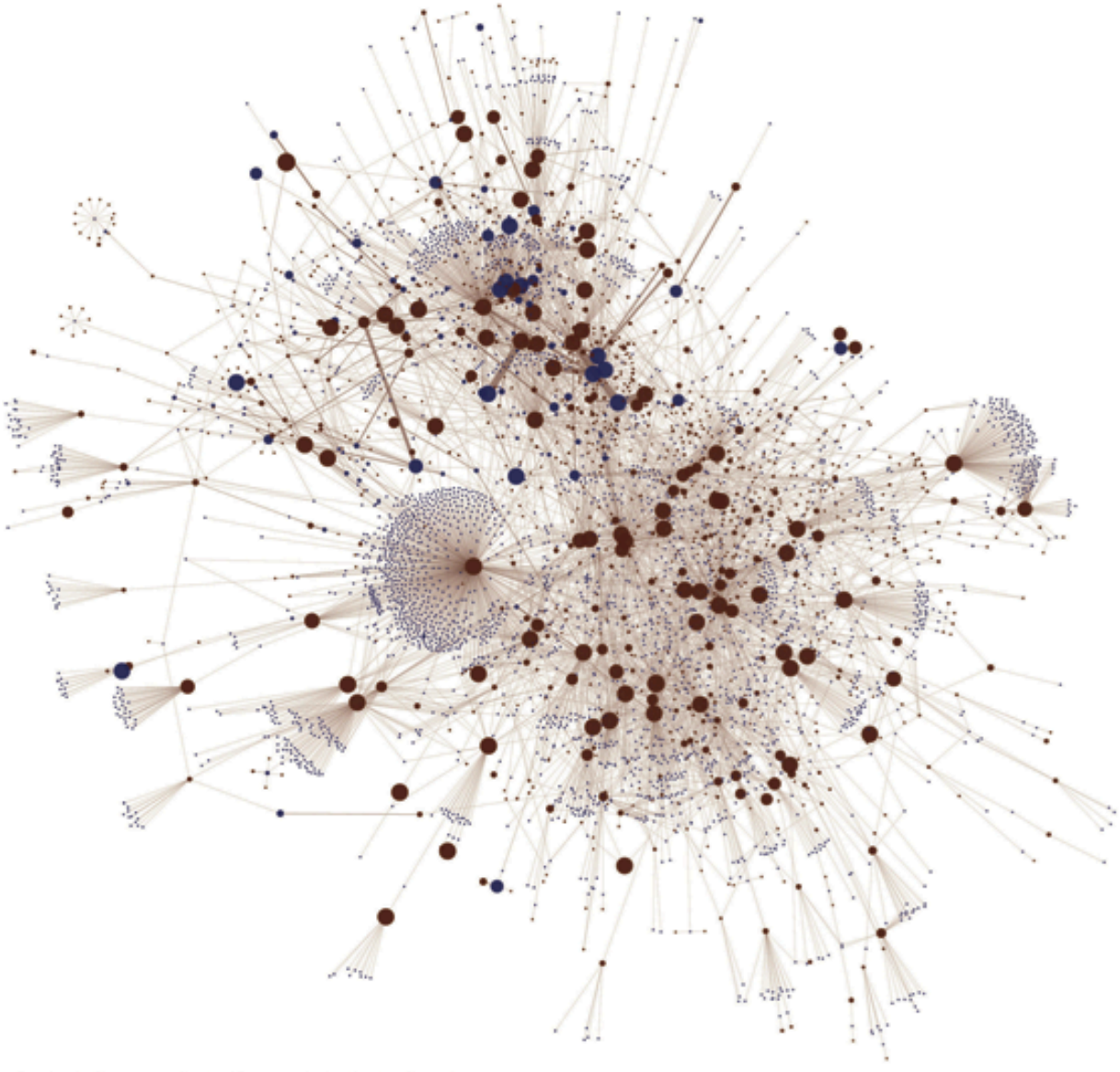

Fig.13: **Log-Log Plot of Church Reference Patterns in Selected Histories**
(the author)

The illustration is explained fully in the text, but can be explained—basically—as a simple plot akin to fig.2, showing the number of references to individual buildings in each text. The points to the high-left represent the most-mentioned buildings in each text. On a linear plot, as in fig.2, these lines would appear less as straight lines, and more as classic exponential curves.