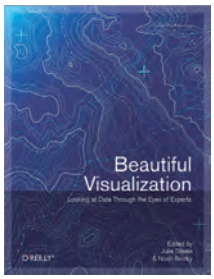


CHAPTER FOURTEEN

The present book chapter was first published in *Beautiful Visualization. Looking at Data through the Eyes of Experts.* Edited by Julie Steele and Noah Iliinsky. Sabastopol, CA: O'Reilly 2010. ISBN: 978-1-4493-7986-5. The full book is available at http://oreilly.com/catalog/0636920000617 Thanks go to O'Reilly for providing a copy of the original for this postprint. You can zoom into the figures at http://revealingmatrices.schich.info
Author contact: maximilian@schich.info

Revealing Matrices

Maximilian Schich

THIS CHAPTER UNCOVERS SOME NONINTUITIVE STRUCTURES in curated databases arising from local activity by the curators as well as from the heterogeneity of the source data. Our example is taken from the fields of art history and archaeology, as these are my trained areas of expertise. However, the findings I present here—namely, that it is possible to visualize the complex structures of databases—can also be demonstrated for many other structured data collections, including biological research databases and massive collaborative efforts such as DBpedia, Freebase, or the Semantic Web. All these data collections share a number of properties, which are not straightforward but are important if we want to make use of the recorded data or if we have to decide where and how our energies and funds should be spent in improving them.

Curated databases in art history and archaeology come in a number of flavors, such as library catalogs and bibliographies, image archives, museum inventories, and more general research databases. All of them can be built on extremely complicated data models, and given enough data, even the most boring examples—however simple they may appear on the surface—can be confusingly complex in any single link relation. The thematic coverage potentially includes all man-made objects: the Library of Congress Classification System, for example, deals with everything from artists and cookbooks to treatises in physics.

As our example, I have picked a dataset that is large enough to be complex, but small enough to examine efficiently. We are going to visualize the so-called Census of Antique Works of Art and Architecture Known in the Renaissance (*http://www.census.de*), which was initiated in 1947 by Richard Krautheimer, Fritz Saxl, and Karl

Lehmann-Hartleben. The CENSUS collects information about ancient monuments—such as Roman sculptures and architecture—appearing in Western Renaissance documents such as sketchbooks, drawings, and guidebooks. We will look at the state of the database at the point just before it was transferred from a graph-based database system (CENSUS 2005) to a more traditional relational database format (CENSUS BBAW) in 2006, allowing for comparison of the historic state with current and future achievements.

The More, the Better?

Having worked with art research databases for over a decade, one of the most intriguing questions for me has always been how to measure the quality of these projects. Databases in the humanities are rarely cited like scholarly articles, so the usual evaluation criteria for publications do not apply. Instead, evaluations mostly focus on a number of superficial criteria such as the adherence to standards, quality of user interfaces, fancy project titles, and use of recent buzzwords in the project description. Regarding content, evaluators are often satisfied with a few basic measures such as looking at the number of records in the database and asking a few questions concerning the subtleties of a handful of particular entries.

The problem with standard definitions such as the CIDOC Conceptual Reference Model (CIDOC-CRM) for data models or the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for data exchange is that they are usually applied *a priori*, providing no information about the quality of the data collected and processed within their frameworks. The same is true of the user interfaces, which give as much indication of the quality of the content as does the aspect ratio of a printed sheet of paper. Furthermore, both data standards and user interfaces change over time, which makes their significance as evaluation criteria even more difficult to judge. As any programmer knows, an algorithm written in the old Fortran language can be just as elegant as and even faster than a modern Python script. As a consequence, we should avoid any form of system patriotism in project evaluation—that is, the users of a particular standard should not have to be afraid of being evaluated by the fans of another.

Even the application of standards we all consider desirable, such as Open Access, is of questionable value: while Open Access provides a positive spin to many current projects, its meaning within the realm of curated databases is not entirely clear. Should we really be satisfied with a complicated but free user interface (cf. Bartsch 2008, fig. 10), or should we prefer a sophisticated API and periodical dumps of the full database (cf. Freebase), which would allow for serious analysis and more advanced scholarly reuse of the data? And if there is Open Access, who is going to pay the salary of a private enterprise data curator?

Ultimately, we must look at the actual content of any given project. As this chapter will demonstrate, when evaluating a database it makes only limited sense to focus on the subtleties of a few particular entries, as usually there is no average information

against which to measure any particular database entry. The omnipresent phenomenon of long tails (Anderson 2006; Newman 2005; Schich et al. 2009, note 5), which we will encounter in almost all the figures in this chapter, suggests that it would be unwise to extrapolate from a few data-rich entries to the whole database—i.e., in the CENSUS, we cannot make inferences about all the other ancient monuments based solely on the Pantheon.

The most neutral of the commonly applied measures remaining for evaluation is the number of records in the database. It is given in almost all project specifications: encyclopedias list the number of articles they contain (cf. Wikipedia); biomedical databases publish the number of compounds, genes, or proteins they contain (cf. Phosphosite 2003–2007 or Flybase 2008); and even search engines traditionally (but ever more reluctantly) provide the number of pages in their indexes (Sullivan 2005). It is therefore no wonder that the CENSUS project also provides some numbers:

> More than 200.000 entries contain pictorial and written documents, locations, persons, concepts of times and styles, events, research literature and illustrations. The monuments registered amount to about 6.500, the entries of monuments to about 12.000 and the entries of documents to 28.000.*

Although these numbers are surely impressive from the point of view of art history, where large exhibition catalogs usually contain a couple of hundred entries, it is easy to disprove the significance of the number of records as a good measure of database quality, if taken in isolation. Just as search engines struggle with near duplicates (cf. Chakrabarti 2003, p. 71), research databases such as the CENSUS aim to normalize data by eliminating apparent redundancies arising from uncertainties in the raw data and the ever-present multiplicity of opinion. Figure 14-1 gives a striking example of this phenomenon. Note that the total number of links remains stable before and after the normalization, pointing to a more meaningful first approximation of quality, using the ratio of the number of links relative to the number of entries: 3/6 vs. 3/4 in this example.

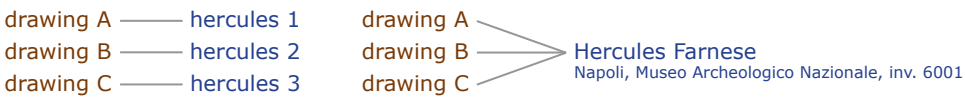


Figure 14-1. *Growing dataset quality by shrinking number of records*

Clearly, more sophisticated measures are required in order to evaluate the quality of a given database. If we really want to know the value of a dataset, we have to look at the global emerging structure, which the commonly used indicators do not reveal. The only thing we can expect in any dataset is that the global structure can be characterized as a nontrivial, complex system. The complexity emerges from local activity (Chua 2005), as the availability of and attention to the source data are highly hetereogeneous

* From *http://www.census.de*, retrieved 9/14/2009.

by nature. Furthermore, every curator has a different idea about the *a priori* data model definitions. As the resulting structural complexity is difficult to predict, we have to measure and visualize it in a meaningful way.

Databases As Networks

Structured data in the fields of art history and archaeology, as in any other field, comes in a variety of formats, such as relational or object-oriented databases, spreadsheets, XML documents, and RDF graphs; semistructured data is found in wikis, PDFs, HTML pages, and (perhaps more than in other fields) on traditional paper. Disregarding the subtleties of all these representational forms, the underlying technical structure usually involves three areas:

- A data model convention, ranging from simple index card separators in a wooden box to complicated ontologies in your favorite representational language
- Data-formatting rules, including display templates such as lenses (Pietriga et al. 2006) or predefined query instructions
- Data-processing rules that act according to the data-formatting instructions

Here, we are interested first and foremost in how the chosen data model convention interrelates with the available data.

As Toby Segaran (2009) pointed out in *Beautiful Data*, there are two ends of the spectrum regarding data model conventions. On one end, the database is amended with new tables, new fields in existing tables, new indices, and new connections between tables each time a new kind of information is taken into consideration, complicating the database model ever further. On the other end, one can build a very basic schema (as shown in Figure 14-2) that can support any type of data, essentially representing the data as a graph instead of a set of tables.

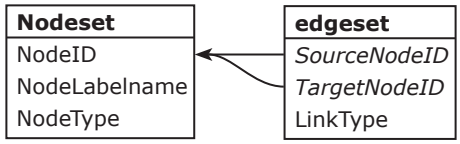


Figure 14-2. *Databases can be mapped to a basic schema of nodes and edges*

Represented in this form, every database can be considered a network. Database entries form the nodes of the network, and database relations figure as the connections between the nodes (the so-called edges or links). If we consider art research databases as networks, a large number of possible node types emerge: the nodes can be the entries representing physical objects such as Monuments and Documents, as well as Persons, Locations, Dates, or Events (cf. Saxl 1947). Any relation between two

nodes—such as "Drawing A was created by Person B"—is a link or edge. Thus, there are a large number of possible link types, based on the relations between the various node types.

A priori definitions of node and link types in the network correspond to traditional data model conventions, allowing for the collection of a large amount of data by a large number of curators. In addition, the network representation enables the direct application of computational analytic methods taken from the science of complex networks, allowing for a holistic overview encompassing all available data. As a consequence, we can uncover hidden structures that go far beyond the state of knowledge at the point of time when the database was conceptualized and that are undiscoverable by regular local queries. This in turn enables us to reach beyond the common measures of quality in our evaluations: we can check how well the data actually fits the data model convention, whether the applied standards are appropriate, and whether it makes sense to connect the database with other sources of data.

Data Model Definition Plus Emergence

To get an idea of the basic structure, the first thing we want to see in a database evaluation is the data model—if possible, including some indicators of how the actual data is distributed within the model. If we're starting from a graph representation of the database, as defined in Figure 14-2, this is a simple task. All we need is a nodeset and an edgeset, which can be easily produced from a relational set of tables; it might even come for free if the database is available in the form of an RDF dump (Freebase 2009) or as Linked Data (Bizer, Heath, and Berners-Lee 2009). From there, we can easily produce a node-link diagram using a graph drawing program such as Cytoscape (Shannon et al. 2003)—an open source application that has its roots in the biological networks scientific community. The resulting diagram, shown in Figure 14-3, depicts the given data model in a similar way as a regular Entity-Relationship (E-R) data structure diagram (Chen 1976), enriched with some quantitative information about the actual data.

The CENSUS data model shown in Figure 14-3 is a metanetwork extracted from the graph database schema according to Figure 14-2: every node type is depicted as a metanode, and every link type is depicted as a metalink connecting two metanodes. The metanode size reflects the number of actual nodes and the metalink line width corresponds to the number of actual links, effectively giving us a first idea about the distribution of data within the database model. Note that both node sizes and link line widths are highly heterogeneous across types, spanning four to five orders of magnitude in our example. Frequent node and link types occur way more often in reality than the majority of less frequent types—a fact that is usually not reflected in traditional E-R data structure diagrams, often leading to lengthy discussions about almost irrelevant areas of particular database models.

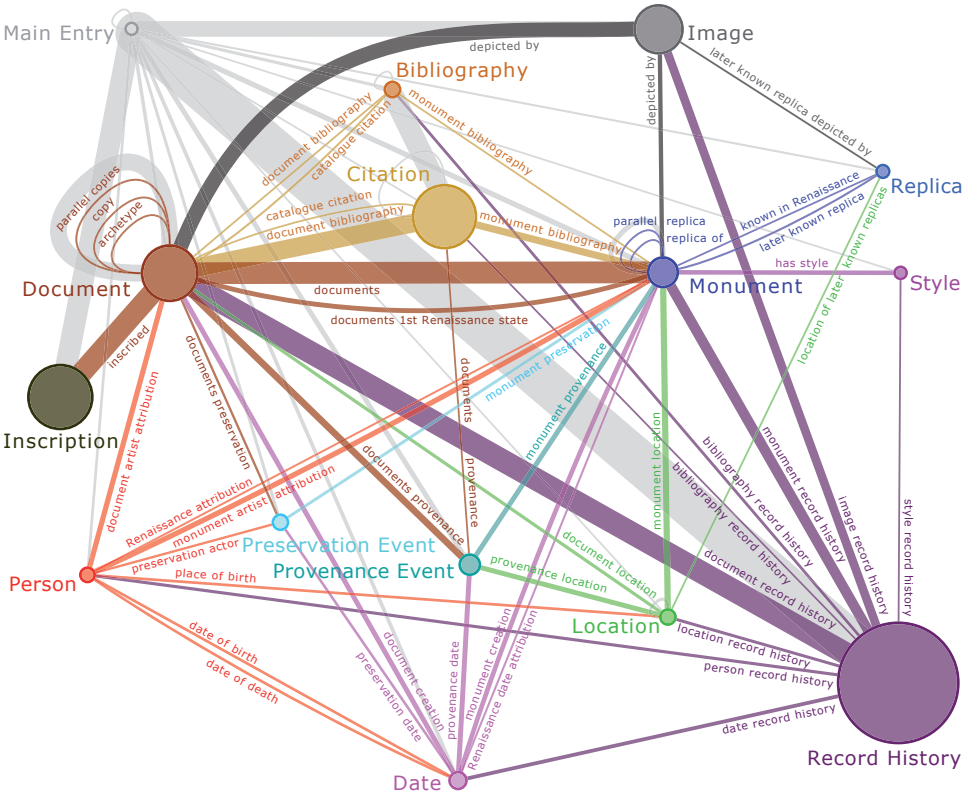


Figure 14-3. *The CENSUS data model as a weighted node-link diagram*

The heterogeneity of node and link type frequency evidenced in Figure 14-3 is not restricted to our example. It is observable in many datasets, including research databases (Schich and Ebert-Schifferer 2009), large bibliographies (Schich et al. 2009), Freebase, and the Linked Data cloud, regardless of whether the number of types is predefined or expandable by the curators. In all cases that I have seen so far, both the number of nodes per node type and the number of links per link type exhibit right-skewed diminishing distributions, which are widely known as *long tails* (Anderson 2006, Newman 2005), and lack a shared average as found in a normal Gaussian distribution. The comparable long-tail structure of hyperlinks in web pages—i.e., of a single link type in only one node type—has been well known for over a decade (*Science* 2009). Figure 14-3 makes clear that the observed heterogeneity is also present at the level of node and link types within more structured data graphs.

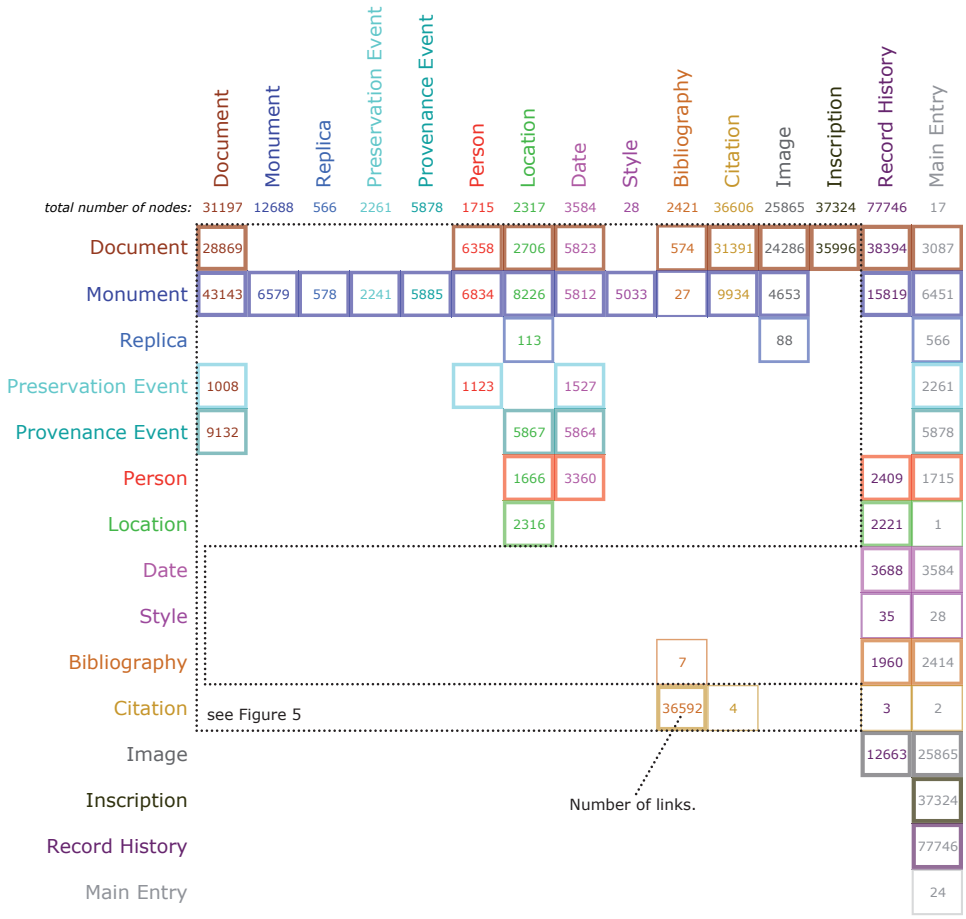


Figure 14-4. *The CENSUS data model as a weighted adjacency matrix*

Network Dimensionality

Looking more closely at Figure 14-3, we can see the central dimensions of the CENSUS database, Monuments and Documents, surrounded by an armature of additional information. Both Monuments and Documents are physical objects, but they differ insofar as the former are the targets and the latter are the sources of the central documentation links. Whereas in general any physical object can function as a Monument or as a Document, the CENSUS divides them into discrete node types because both groups belong to different periods (Classical Antiquity and Western Renaissance): ancient Roman sculptures and architecture as documented by Renaissance drawings, sketchbooks, text, etc.

In addition to these central dimensions, there is another node type representing physical objects called Replica, used for later-known replica Monuments that were discovered only after the defined Renaissance time frame. If the CENSUS is to be generalized to encompass the entire time frame from Antiquity until today, it would make sense to combine Monuments, Documents, and Replicas into a single physical object node type, as all functions are defined by the presence of certain links pointing into or out of a particular node. In the early 1980s, when the data model was initially conceived, its design was influenced by certain functionality constraints regarding relational databases. These constraints no longer apply, so such a change is now possible.

Distributed around the physical objects in Figure 14-3, we find Persons, Locations, and time ranges (such as Date and Style). Relations between all these dimensions are mostly modeled using direct links. For example, each Person is connected directly to a place of birth and a date of birth, making it impossible to disambiguate two alleged Birth Events (such as Venice 1573 and Bologna 1568) in a single Person without further comment.

Other example shortcuts include the document artist attribution and the 1st Renaissance state documentation. Again disambiguation is impossible without further comment. Regarding artist attribution, the CENSUS curators are guided to make a decision instead of recording multiple opinions. In the case of 1st Renaissance state documentation, there is only a single instance by definition. Further states are documented as Preservation Events—an obvious opportunity to simplify the data model.

Preservation and Provenance Events are a notable exception to the aforementioned shortcuts. They state that a particular Monument was altered by a Person or present at a particular Location, at a particular Date, as documented by a particular Document. Both Preservation and Provenance Events allow for easy disambiguation.

Differing opinions across Documents can be reflected by multiple Events, gluing together the respective Monuments, Persons, Locations, and Dates. As with physical objects, the nature of the Events is defined by the presence of certain links. As a consequence, the data model could be generalized further, as was done in projects inspired by the CENSUS such as the Winckelmann Corpus (2000). In general, Events boil down to so-called star motifs (cf. Milo et al. 2002) with a particular combination of link types. Today, Event-like constructions are a standard feature of many database models, such as Freebase, where they are called *compound value types*. In principle, we could also look for such Events in other networks with typed links, where they are not consciously explicit but rather inherent as emergent star motifs (as in the Linked Data graph).

The CENSUS becomes an authoritative—i.e., citable—source of information by providing a variety of metadimensions, such as the (modern) Bibliography. The Bibliography is subdivided into Citations, which are in turn represented as a separate node type. Another source dimension is the Image node type, which contains photographs taken from major photo libraries. Again, both the Bibliography and the Images represent functions of physical objects, which are defined by their adjacent links.

The remaining node types include the Record History, where curators log their actions on other nodes, and the Main Entry dimension, which was probably dissolved during the conversion of the CENSUS to a relational database. In the former graph-based system, due to the lack of tables the Main Entries figured as database chapters, facilitating navigation by bundling together all Persons, Locations, etc.

The Matrix Macroscope

The node-link diagram in Figure 14-3 is only one possibility for depicting the CENSUS data model. As with any network consisting of nodes and links, we can also depict it in the form of a so-called *adjacency matrix* (cf. Garner 1963; Bertin 1981; Bertin 2001; Henry 2008), as shown in Figure 14-4. Here, the node types are represented as the vertical columns and horizontal rows of a table, with link information appearing in the cells. Regarding the place of birth, for example, you can imagine the link pointing from the Person row into the Location column across the respective cell.

As in the node-link diagram, it is also possible to depict the total number of links occurring between two node types in the adjacency matrix; in place of the line width in Figure 14-3, now the explicit number appears in the relevant cell. This highlights the main difference in switching the representation to a matrix: our attention now focuses on the links, rather than on the nodes. It is striking that the matrix in Figure 14-4 not only shows the connections between node types, but also makes immediately clear which node types are not directly connected. In other words, the matrix indicates positive as well as negative correlation. One example of this is the absence of links from the Bibliography node type to authors, publication locations, and publication dates; though the CENSUS provides this information, it is only implicit in the node description text and node label abbreviations (e.g., "Nesselrath 1993"). Of course, we can also spot this absence of information in the node-link diagram, but the matrix makes it way more obvious.

Going beyond the total number of links between two node types, we can put a variety of other useful information into the matrix cells. In Figure 14-5, for example, we see a node-link diagram of all the nodes and links occurring between two node types in a cell. We generate such a diagram using a layout algorithm (such as the yFiles organic layout algorithm, part of the Cytoscape application), which is relatively inexpensive from a computational point of view. As a consequence, all of the explicit node-link data in the database appears in the data model matrix.

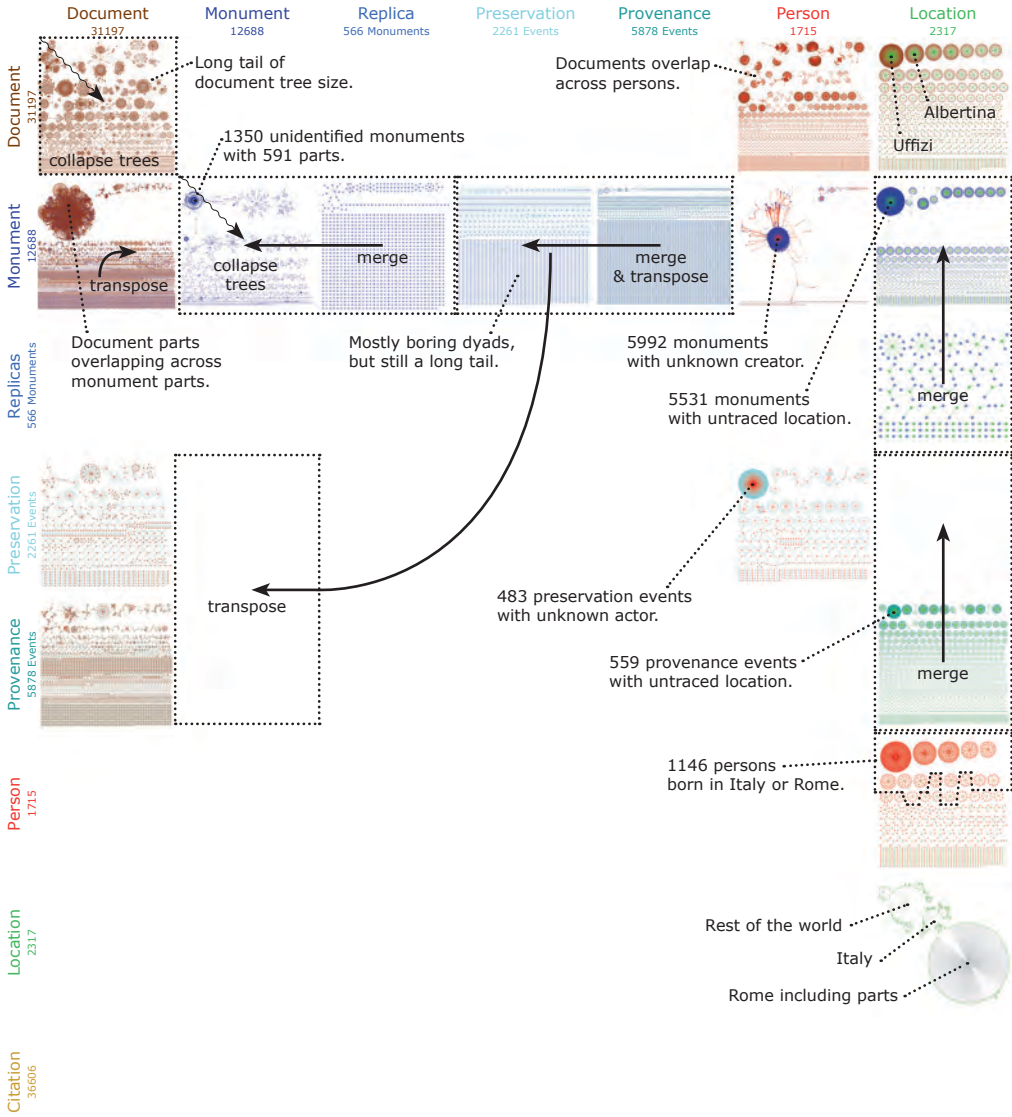
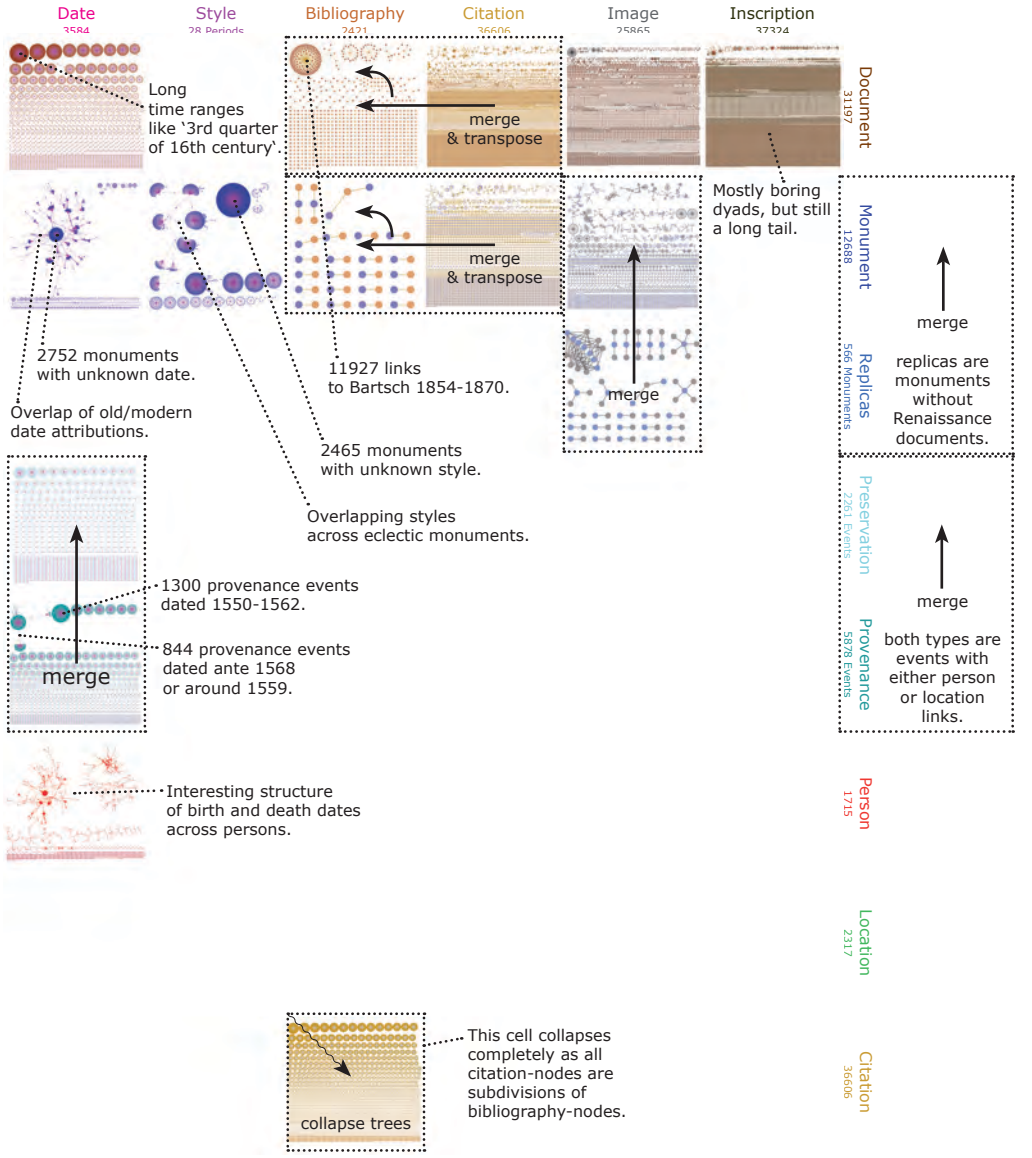


Figure 14-5. *The CENSUS data model as an adjacency matrix, enriched with node-link diagrams, i.e., actual data*


Date
Style
Bibliography
Citation
Image
Inscription
Document
31197
Long
time ranges
like '3rd quarter
of 16th century'.
merge
& transpose
Mostly boring
dyads, but still
a long tail.
Monument
12688
merge
& transpose
merge
2752 monuments
with unknown date.
Overlap of old/modern
date attributions.
11927 links
to Bartsch 1854-1870.
Replicas
566 Monuments
merge
replicas are
monuments
without
Renaissance
documents.
2465 monuments
with unknown style.
Overlapping styles
across eclectic monuments.
Preservation
2261 Events
1300 provenance events
dated 1550-1562.
merge
844 provenance events
dated ante 1568
or around 1559.
Provenance
5878 Events
merge
both types are
events with
either person
or location
links.
Interesting structure
of birth and death dates
across persons.
Person
1715
Location
2317
collapse trees
This cell collapses
completely as all
citation-nodes are
subdivisions of
bibliography-nodes.
Citation
36606

Looking at the result in Figure 14-5, we can learn a lot about the database. At first glance, we can see that there are a few cells in which the structure looks more complex, whereas in the majority of cells we find a rather boring collection of stars and some dyads connecting two nodes exclusively. Another thing we can see is that all of the cells contain disconnected networks, in the sense that they are split into discrete components (i.e., groups of connected nodes). It is intriguing that here again we do not find a widespread average for component size. Wherever we look, we see a long tail. A prominent example is the Document-Location cell, wherein we see a clearly diminishing sequence of stars, connecting ever fewer Documents to single Locations; but even in the flattest cases, such as in the Document-Image cell, we find a few larger connected groups, followed by a huge amount of dyads.

A more diluted form of long tail is found in the Location-Location cell. It contains a hierarchy of geographical places rooted in a node representing the world, subdivided into countries, regions, and towns, down to individual collections. The number of subdivisions per Location is again distributed in a heterogeneous way. The majority of subdivisions are found within the country of Italy, almost eclipsing the rest of the world. The most prominent Location is unsurprisingly the city of Rome, which is subdivided into numerous collections. Its prominence reminds me of the oversized space dedicated to the hands in the sensomotory *homunculus* of the human brain (Penfield and Rasmussen 1950; Dawkins 2005, pp. 243–244)—the CENSUS seems to have a *romunculus*. Just as an overly large area of our brain's motor cortex is dedicated to hand–eye coordination and the sense of touch in our hands, the CENSUS Location hierarchy seems to be biased toward sculpture collections in Rome. Like a master pianist, whose centers for dexterity and manual control occupy even more space in the cortex than they would in a regular person, the CENSUS seems to be defined by specialization—such as the addition of Ulisse Aldroandi's famous books (1556 and 1562), which list thousands of sculptures in Roman collections (cf. Schich 2009, pp. 124–125).

Another interesting feature of Figure 14-5 is the disproportionally large stars found in a number of cells. Some of the stars are natural properties of the data, as in the case of the 11,927 Document nodes linked to the Bibliographic node Bartsch 1854–1870, or the 1,146 Persons born in Italy or Rome. However, most of the giant stars are artifacts related to unknown entries, such as an unidentified Monument, unknown Person, untraced Location, unknown Date, or unknown Style; all of these single nodes connect confirmed gaps of information in order to facilitate their further curation. There are 1,350 unidentified Monuments, 5,992 Monuments with unknown creators, 5,531

Monuments with untraced Locations, 2,752 Monuments with unknown Dates, 2,465 Monuments with unknown Styles, 483 Preservation Events with unknown actors, and 559 Provenance Events with untraced Locations in our dataset. To be sure, the presence of all these unknown entries is not an error; the attribution of an unknown Date could, for example, refute an incorrect Renaissance date attribution. However, the numbers provide a feeling of how incomplete our knowledge is. Another consideration is that, if we want to analyze the network structure of each cell, we have to break (or *denormalize*) the unknown nodes; otherwise, the untraced Location shortcut node would, for example, connect many unrelated nodes located at many different unknown places.

Reducing for Complexity

If we look back at Figure 14-3 for a moment, we can see that there are 31,197 Document records in the CENSUS database, of which only 3,087 are connected to the document authority under Main Entry. This points to an important fact: large Documents in the database are represented as trees of nodes. There are in fact only 3,087 Documents, including 28,110 subordinate nodes representing pages, figures, and quadrants within those figures or paragraphs of text—a fact until now rarely communicated about this database. The same is true for Monuments: here again, a small percentage of the records—in particular, the Architecture category—is subdivided into trees of nodes including building parts, rooms, and even tiny individual features of architectural decoration. A third example is the Bibliography, which is subdivided into Citations, such as paragraphs of text in modern scholarly books.

The consequence of all these subdivisions in Figure 14-5 is that particular links point from and to particular subnodes: from Monument parts to Document parts instead of from entire Monuments to entire Documents, or from a feature of a decorated column base to a particular quadrant in a sketchbook figure. The function of all these subdivisions is to enable data storage without a significant loss of information. However, the questions we can resolve in this configuration are often too specific. In order to uncover more interesting global properties of the data and answer questions such as how many sketchbooks a group of Monuments appears in (not how many figures there are in general), or how often they are cited in books (not how many citations there are in general), we have to refine the matrix. A solution for this problem is to collapse the subdivided Documents, Monuments, and Bibliographic Citation nodes as shown in Figure 14-6 and redraw the entire matrix as in Figure 14-7(a).

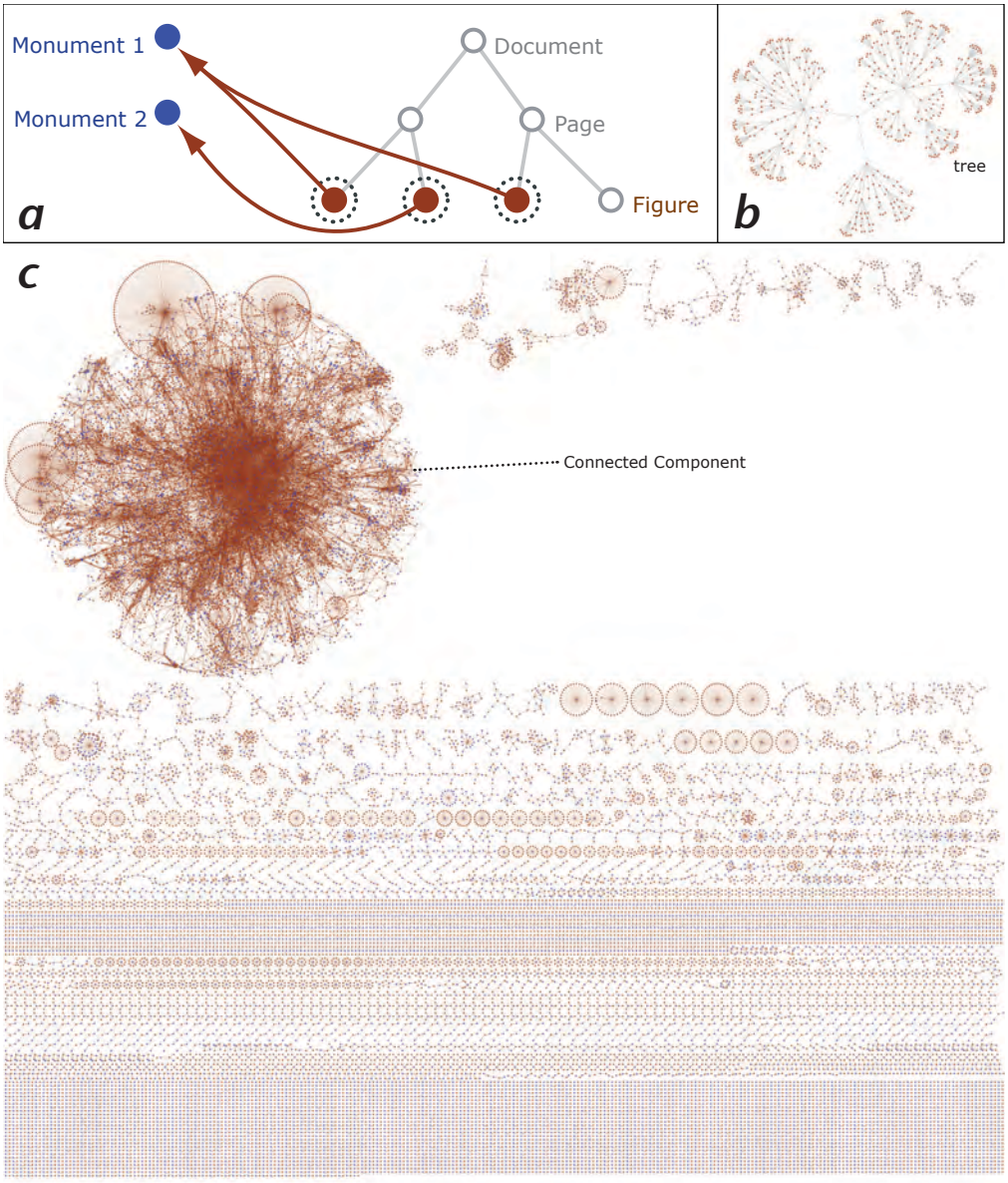
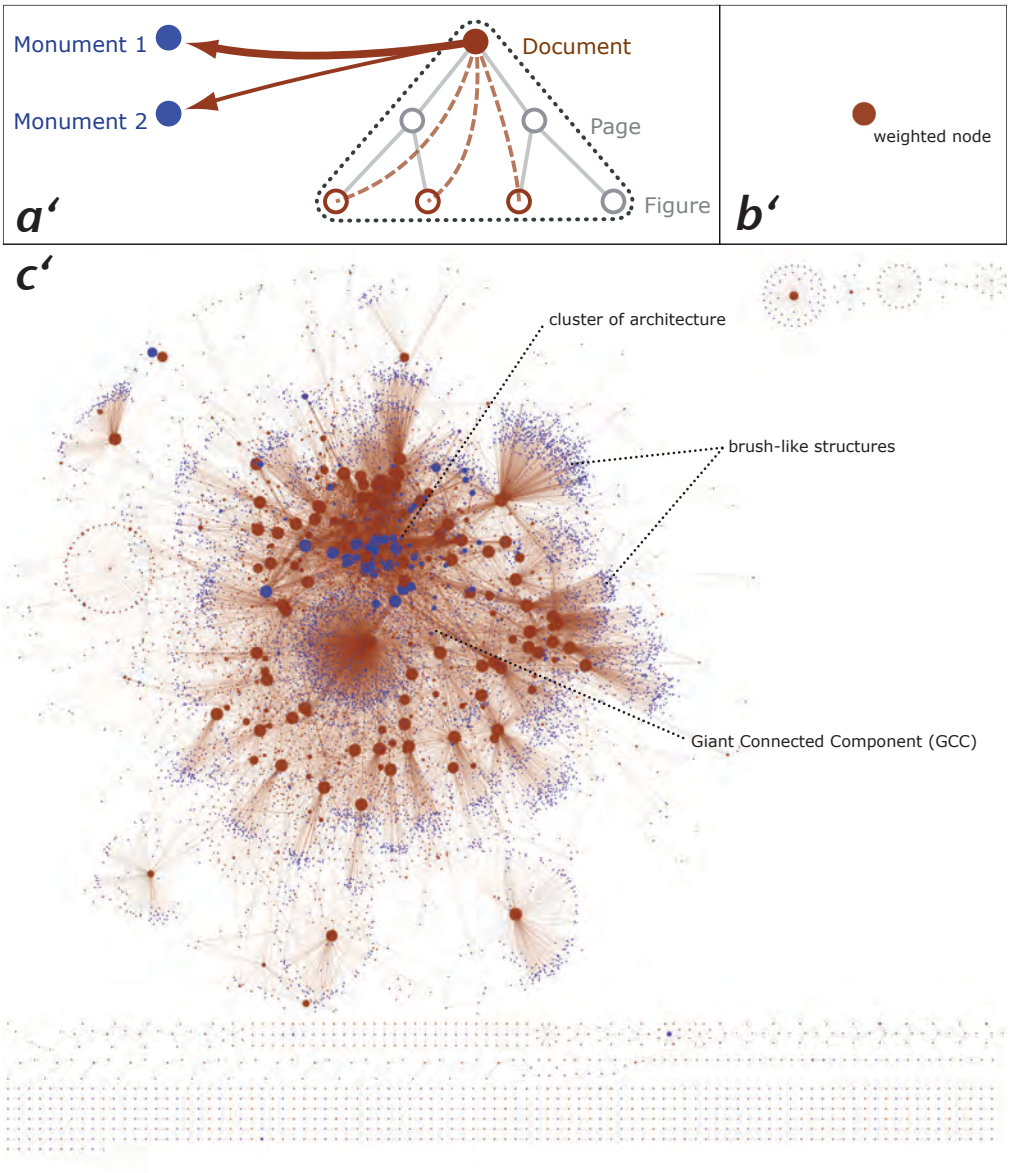


Figure 14-6. *Collapsing subdivided entries in the raw data uncovers interesting complex features*


Monument 1
Monument 2
Document
Page
Figure
a‘
weighted node
b‘
c‘
cluster of architecture
brush-like structures
Giant Connected Component (GCC)

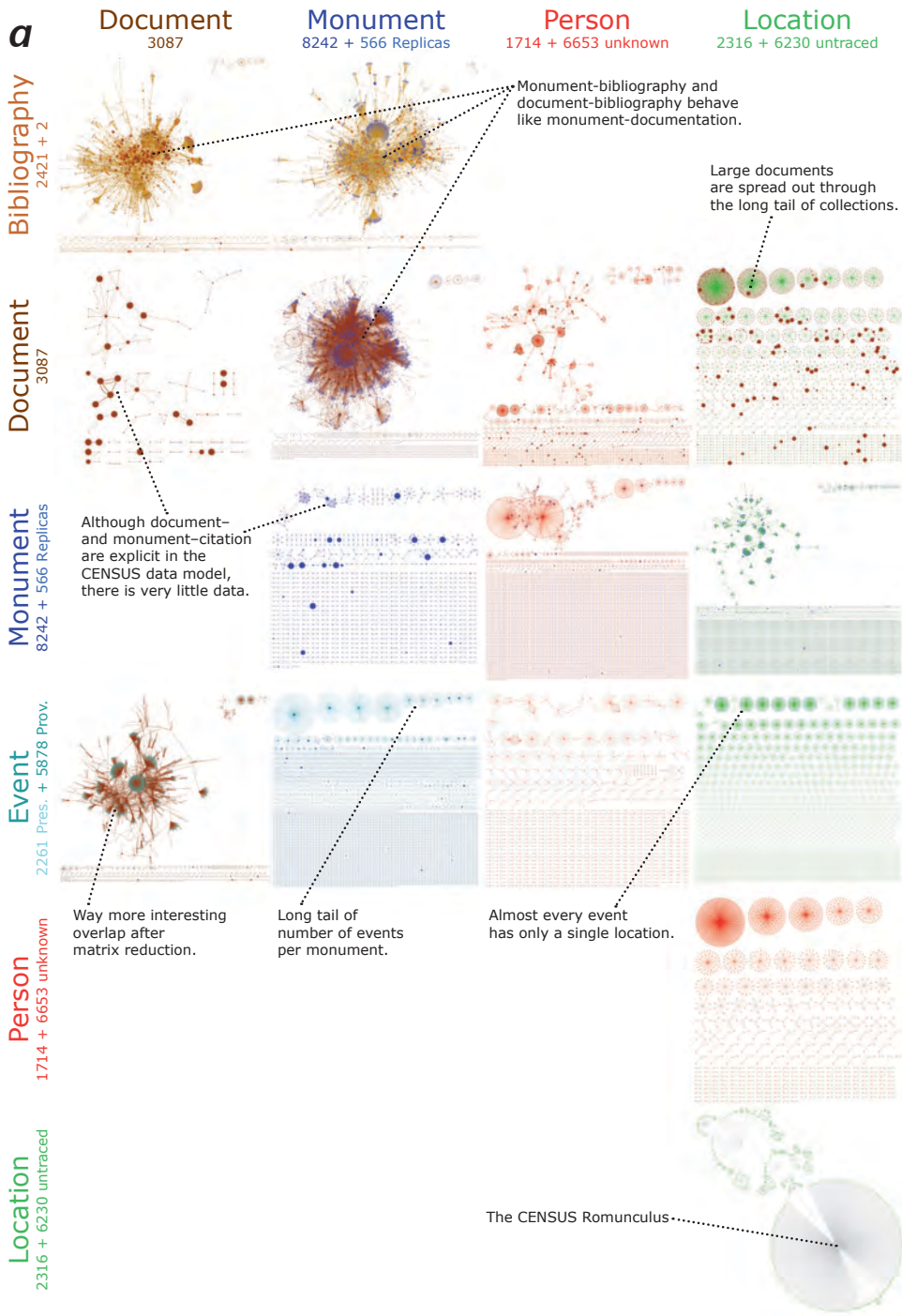
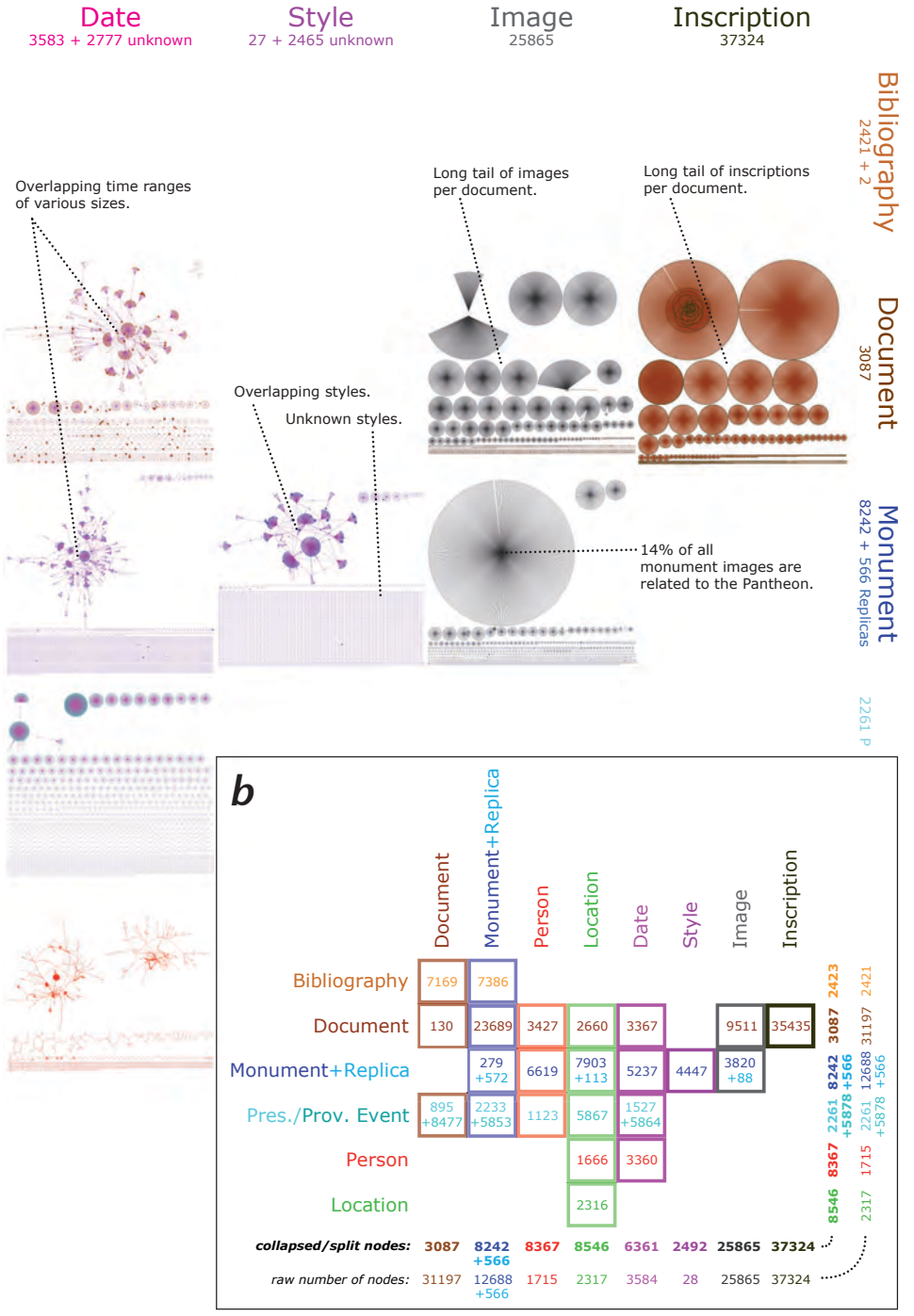


Figure 14-7. *The refined CENSUS data model matrix, enriched with node-link diagrams (a), and in the basic weighted form (b)*


Date
3583 + 2777 unknown
Style
27 + 2465 unknown
Image
25865
Inscription
37324
Bibliography
2421 + 2
Document
3087
Monument
8242 + 566 Replicas
Overlapping time ranges of various sizes.
Long tail of images per document.
Long tail of inscriptions per document.
Overlapping styles.
Unknown styles.
14% of all monument images are related to the Pantheon.
b
Document
Monument+Replica
Person
Location
Date
Style
Image
Inscription
Bibliography
Document
Monument+Replica
Pres./Prov. Event
Person
Location
7169
7386
130
23689
3427
2660
3367
9511
35435
279 +572
6619
7903 +113
5237
4447
3820 +88
895 +8477
2233 +5853
1123
5867
1527 +5864
1666
3360
2316
collapsed/split nodes:
3087
8242 +566
8367
8546
6361
2492
25865
37324
raw number of nodes:
31197
12688 +566
1715
2317
3584
28
25865
37324

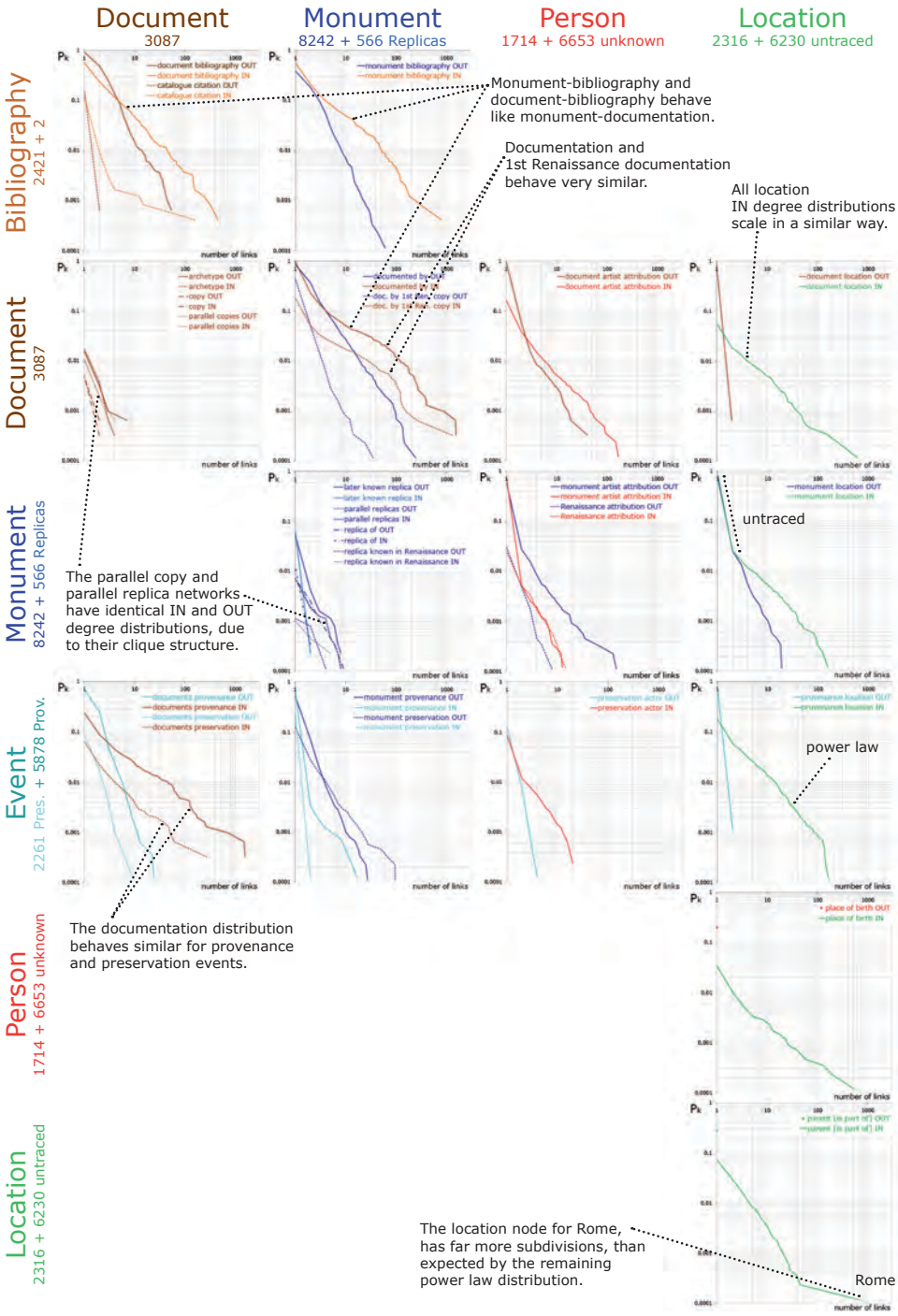


Figure 14-8. *The refined CENSUS data model matrix, enriched with degree distribution plots*

Date
3583 + 2777 unknown

Style
27 + 2465 unknown

Image
25865

Inscription
37324

Bibliography
2421 + 2

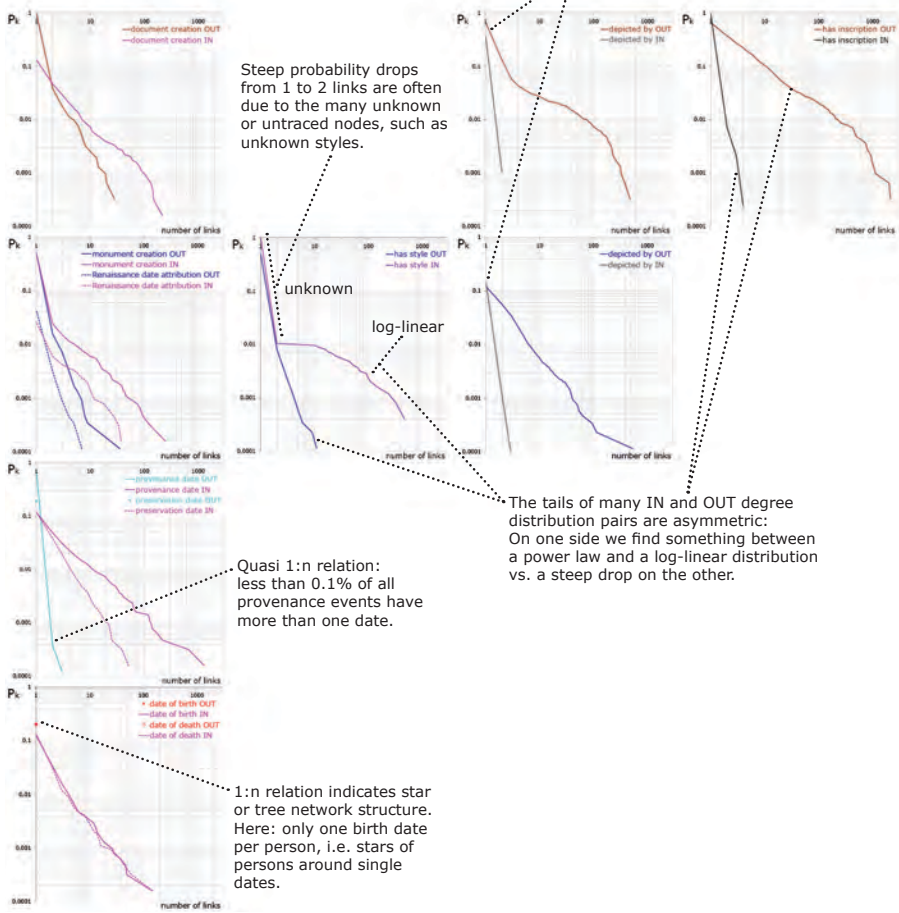
Document
3087

Monument
8242 + 566 Replicas

Event
2261 Pres. + 5878 Prov.

Person
1714 + 6653 unknown

Location
2316 + 6230 untraced



Collapsing the Documents, Monuments, and Bibliographic Citation trees to single nodes works as follows (cf. Schich 2009, p. 28–37). In Figure 14-6(a), we see a raw Document tree: a book, with pages, which in turn are subdivided into figures. Single links point to multiple Monuments or Monument parts. In order to collapse the tree, we represent the book as a single node and combine all the links adjacent to the subdivisions, as shown in Figure 14-6(a′). To preserve as much information as possible, we assign a weight to the new node reflecting the number of collapsed subdivisions and another weight to the links, signifying the number of occurrences in the book. Graphically, our weights now correspond to the node size and the line width: the larger the node of a book is, the more subnodes it contains in its collapsed tree; the broader a line is, the more links it represents. Exemplified in real data, every Document tree in the Document-Document cell of the raw matrix will be reduced to a single node, as shown in Figures 14-6(b)/(b′). Matrix cells that look boring or simple in the raw state become more complex and interesting after the collapse, as in the case of the Document-Monument cell enlarged in Figures 14-6(c)/(c′).

The most striking feature of the refined cell in Figure 14-6(c′) is the emergence of a so-called Giant Connected Component (GCC), which connects almost 90% of all Monuments and Documents in the CENSUS—a phase transition phenomenon known from many other complex networks and bearing many important implications regarding the propagation of information (Newman, Barabási, and Watts 2006, pp. 415–417; Schich 2009, pp. 171–172). In the core of the GCC, we can see a cluster of large architectural Monuments, which are connected to large overview Documents, such as guidebooks, sketchbooks, and city maps. A surprising feature in the periphery of the GCC is the dominance of brushlike structures connected to large Document nodes: obviously a large percentage of all Monuments in the CENSUS are connected to only one single Document, either because the Documents lack sufficient information or because (for whatever reason) the curators did not identify and normalize them.

As the Document, Monument, and Bibliography trees are collapsed, the consequences affect the whole matrix. Effectively, the diagonal Document-Document and Monument-Monument cells are thinned out, leaving only a few interesting links, such as archetype citation and parallel copy relations. The Citation-Bibliography cell collapses completely.

Further Matrix Operations

Beyond breaking unknown nodes and collapsing the trees of subdivisions, we can do a number of other operations on the raw matrix in Figure 14-5. As with any adjacency matrix, we can sort (or *permutate*) the columns and lines along the horizontal and vertical axes, without losing any information (Bertin 1981; Bertin 2001). We can also transpose cells such as Monument-Event to Event-Monument, or even the whole Bibliography column to a Bibliography line, effectively reversing the direction of the links. Finally, we can merge equivalent node types—such as Provenance

and Preservation Events, Monuments and Replicas, or Bibliography and Citation—by creating node supertypes, such as Event, Monument, and Bibliography. The merge reduces the number of columns and lines in the matrix and allows each cell to occupy more space in the visualization. Beyond that, the literature on matrix visualization contains many more possible operations (cf. Henry 2008).

The Refined Matrix

Figures 14-7(a) and (b) show the final result of all refining operations discussed so far. The whole matrix is now more concise, clear, and informative. We can easily see how the CENSUS data is distributed within the data model: Monument- and Document-Bibliography obviously behave like Monument-Documentation, exhibiting a wealth of data. For Document-Document and Monument-Monument dependency relations (such as citations), on the other hand, there is hardly any data, even though the respective links are explicit in the data model. Apparently the data curation workflow was not set up in the right way to collect this kind of information systematically.

As in the raw matrix, we find a long tail of component sizes in every refined cell. Some of the cells still contain mostly stars, as is true for the number of Events per Monument, Images per Document/Monument, Inscriptions per Document, or Events per Location. An interesting case involves the Document-Location cell, where we can see that large Documents are spread throughout all sizes of collections, from the Uffizi in Florence to individual private collections owning a single sketchbook.

Other cells show more overlapping structures, as is the case with overlapping Dates (or time ranges) across Documents and Monuments, or Styles across a few eclectic Monuments such as the Arch of Constantine, which brings together reliefs from different periods of the Roman Empire. Unsurprisingly, Monument-Documentation and the related Bibliography contain the most complex overlap, as this is the central focus of the CENSUS project.

Scaling Up

Readers involved in the network field may point out that the use of node-link diagrams in the matrix, as seen in Figure 14-7(a), is not feasible for datasets an order of magnitude larger than the CENSUS, let alone as large as the entire Semantic Web. Indeed this is a problem, so the question is how to scale the presented approach to really large databases. One solution is to use degree distribution plots or even more sophisticated numerical network measures to get an idea about the actual data within the data model.

In Figure 14-8, we plot a cumulative IN- and OUT-degree distribution (Broder et al. 2000; Newman 2005) for every link type occurring in a matrix cell. As every link points OUT of the source node type and IN to a target node type, there are two distributions for every link type in each cell. The *x*-axis of each plot indicates the number of

links, *k*; the *y*-axis provides the cumulative probability, P(*k*), that a node has at least *k* links. Note that the distributions are plotted on a log-log scale, meaning that the tick marks indicate a rapid decay from 100% to 0.01% on the *y*-axis and a rapid increase from 1 to 3,000 on the *x*-axis. (In a regular linear projection, the slope of each distribution would be so steep that we would not see anything interesting.) It is striking that there is not a single Gaussian bell curve in the plots, as we would expect for, say, the average heights of people. Instead, we find a whole zoology of long tails ranging from beautiful power-laws to log-linear curves, with less clean, bumpier distributions in between.

Nearly all IN and OUT distribution pairs appear to be asymmetric. Birth Dates, for example, are connected to Persons in a 1:*n* manner, where *n* is highly heterogeneous. This is no surprise, as this area of information is not subject to the multiplicity of opinion, as we would expect in a prosopographic database, which would focus on people instead of objects. Other areas, such as the occurrence of Locations in Provenance Events, exhibit a quasi 1:*n* constraint, as it is highly improbable but not impossible for an event to involve more than one location. The most interesting asymmetry is found in true *n*:*n* relations, such as the central Monument-Documentation link, where we find distributions with different slopes on both sides of the link. Right now, it is not entirely clear how this asymmetry can be fully explained; however, by comparing a number of data sources, it becomes apparent that the different shapes of these distributions are caused by a variety of factors, such as physical restrictions and accessibility of the source data, as well as attention and other cognitive limits on the side of the curators.

The only symmetric link relationship in the CENSUS can be found in the parallel copy and parallel replica links in the Document-Document and Monument-Monument cells, respectively. Ideally, the IN- and OUT-degree distributions should be identical, as the relevant nodes are fully connected to so-called "cliques." In reality, both link types become more asymmetric the further down we go into the tail of the distributions, as large cliques are hard to maintain. As I recommended to the CENSUS project in 2003, it makes more sense to connect to an unknown archetype Document with *n* links than to manually connect *n* parallel copies with *n* * (*n*–1) links amongst each other.

Similarly, the behavior of certain relationships that we spotted in Figure 14-7, such as the equivalence of Monument-Bibliography and Monument-Documentation, is confirmed in Figure 14-8 (cf. Schich and Barabási 2009). Not only is there an obvious similarity between these cells, but the same functional equivalence is found in different link types in a single cell. A convincing example are the almost parallel distribution slopes of general documentation and 1st Renaissance documentation in the Document-Monument cell; the same is true for provenance and preservation documentation in the Event-Documentation cell. The Location IN degree scales in a very similar way across all relevant cells in the Location column. Two exceptions to this

observed regularity are the steep drop of probability from one to two Monuments per Location (due to the many untraced Monuments) and the accelerating tail in the Location-Location cell (caused by the *romunculus* phenomenon).

One last thing we can observe in all the plots is the fraction of nodes per node type, which are inherent in the individual networks constituted by each individual link type. Looking at the value where the respective curve crosses the *y*-axis shows us, for example, that less than 15% of all Images are connected to Monuments, and less than 40% to Documents. Inversely, we can conclude that at least 45% of the 24,000 images scanned by the CENSUS project's publishing partner in 1994 were still not linked in the database in 2005.

Further Applications

The visualizations presented here can serve as a starting point for a variety of activities. Besides the evaluation of particular project goals by funders and project leaders, further areas of study include the identification of interesting research topics: every single cell in the matrix could be the subject of an extensive investigation, as illustrated in my PhD dissertation, which deals with monument documentation and visual document citation (Schich 2009). Multiple cells that promise an interesting interplay could also be combined within such a study—for example, in order to build trajectories of objects and persons involved in a variety of events across time and space (cf. González, Hidalgo, and Barabási 2008), or to study the effects of network interaction (Leicht and D'Souza 2009). Finally, a number of equivalent visualizations could be used to compare entire databases that already use similar data models, such as the Winckelmann Corpus and the CENSUS, or databases that can be mapped to the same standard, such as the CIDOC CRM.

Instead of dissecting the databases in the way discussed here, it might also be interesting to combine separate networks in a similar visualization. Candidates for such a combination can easily be found in the multipartite universe of conceivable networks (for example, citation, coauthorship, and image-tagging databases in the social sciences, or gene-transcription, protein-protein interaction, and gene-disease databases in biology).

The coarse graining we obtained by collapsing the Document, Monument, and Bibliography trees can also be achieved in many other ways; for example, by concentrating on particular subtrees, or with more sophisticated methods such as block-modelling (cf. Wassermann and Faust 1999, pp. 394–424) or community finding (cf. Lancichinetti and Fortunato 2009; Ahn, Bagrow, and Lehmann 2009), practically addressing the question of how nodes and links in a network are actually defined (cf. Butts 2009).

Finally, the presented combination of matrix and node-link diagrams can be expanded; for example, by placing node-link/matrix combinations (Henry, Fekete, and McGuffin 2007) or scalable image matrices (Schich, Lehmann, and Park 2008) in relevant cells of the data model matrix.

Conclusion

As this chapter has illustrated, enriched and refined data model matrices are very useful for database project evaluation, exposing many nonintuitive data properties that are hard to uncover by simply using the database or looking at the commonly used indicators of quality. As data becomes more accessible in the form of Linked Data, RDF graphs, or open dumps of relational tables, the presented methods can be applied by funders or the projects themselves, within a very short time frame in a mostly automated process.

The visualizations shown here present the first comprehensive big picture of the entire CENSUS database, where we can see the initial data model definition as well as the emerging complex structure in the collected data. By looking at the visualizations, we found out that many of the numbers given in the project description were incomplete or even misleading. Some of the new numbers may be smaller than the initially presented ones, but as we have learned from our analysis, sometimes a little less is more—and more is different (Anderson 1972).

Acknowledgments

For their useful feedback, I would like to thank my audiences at NetSci09 in Venice and SciFoo09 in Mountain View, as well as my colleagues at the BarabásiLab at Northeastern University in Boston. Further thanks go to Ralf Biering and Vinzenz Brinkmann of Stiftung Archäologie in Munich for providing the data and the German Research Foundation (DFG) for funding my research. For a comprehensive bibliography regarding the CENSUS database see Schich 2009, p. 13, notes 20–25.

References

The presented visualizations are available online in large resolution at *http://revealingmatrices.schich.info*.

Ahn, Yong-Yeol, James P. Bagrow, and Sune Lehmann. 2009. "Link communities reveal multi-scale complexity in networks." *http://arxiv.org/abs/0903.3178v2*.

Aldroandi, Ulisse. 1556/1562. "Appresso tutte le statue antiche, che in Roma in diversi luoghi, e case particolari si veggono, raccolte e descritte (...) in questa quarta impressione ricorretta." *Le antichità della città di Roma*. Ed. Lucio Mauro. Venice.

Anderson, Chris. 2006. *The Long Tail*. New York: Hyperion. *http://www.thelongtail.com*.

Anderson, P.W. 1972. "More is different." *Science* 177, no. 4047: 393–396.

Bartsch, Adam. 1854–1870. *Le Peintre-Graveur, nouvelle edition*. v. 1–21. Leipzig: Barth.

Bartsch, Tatjana. 2008. "Distinctae per locos schedulae non agglutinatae" – Das Census-Datenmodell und seine Vorgänger. *Pegasus* 10: 223–260.

Bertin, Jaques. 1981. *Graphics and Graphic Information Processing*. Berlin: de Gruyter.

Bertin, Jacques. 2001. "Matrix theory of graphics." *Information Design Journal* 10, no. 1: 5–19. doi: 10.1075/idj.10.1.04ber.

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2009. "Linked data—The story so Far." *International Journal on Semantic Web & Information Systems* 5, no. 3: 1–22.

Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. "Graph structure in the Web." *Computer Networks* 33, no. 1–6: 309–319. doi:10.1016/j.physletb.2003.10.071. [erratum: doi:10.1016/S1389-1286(00)00083-9]

Butts, Carter. 2009. "Revisiting the foundations of network analysis." *Science* 325, no. 5939: 414–416. doi: 10.1126/science.1171022.

CENSUS. 1997–2005. *Census of Antique Works of Art and Architecture Known in the Renaissance*. Ed. A. Nesselrath. Munich: Verlag Biering & Brinkmann/Stiftung Archäologie. *http://www.dyabola.de*.

CENSUS BBAW. 2006. *Census of Antique Works of Art and Architecture Known in the Renaissance*. Ed. Berlin-Brandenburgische Akademie der Wissenschaften and Humboldt-Universität zu Berlin. *http://www.census.de*.

Chakrabarti, Suomen. 2003. *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco, CA: Morgan Kaufmann.

Chen, Peter P.S. 1976. "The entity-relationship model—Toward a unified view of data." *ACM Transactions on Database Systems* 1, no.1: 1–36. doi: 10.1145/320434.320440.

Chua, Leon O. 2005. "Local activity is the origin of complexity." *International Journal of Bifurcation and Chaos* 15: 3435–3456. doi: 10.1142/S0218127405014337.

Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff, eds. 2006. *Definition of the CIDOC Conceptual Reference Model (CIDOC-CRM), Version 4.2.1. http://www.cidoc-crm.org/docs/cidoc_crm_version_4.2.1.pdf*.

Dawkins, Richard. 2005. *The Ancestor's Tale. A Pilgrimage to the Dawn of Life*. London: Phoenix.

DBpedia. 2009. *DBpedia*. Sören Auer, Christian Bizer, and Kingsley Idehen, admins. Leipzig: Universität Leipizig; Berlin: Freie Universität Berlin; Burlington, MA: OpenLink Software. *http://dbpedia.org*.

Doreian, P., V. Batagelj, and A. Ferligoj. 2005. *Generalized Blockmodeling (Structural Analysis in the Social Sciences)*. Cambridge: Cambridge University Press.

Flybase. 2008. Rachel Drysdale and the FlyBase Consortium. FlyBase. *Drosophila*: 45–59. doi: 10.1007/978-1-59745-583-1_3. See also *http://flybase.org/static_pages/docs/release_notes.html*.

Freebase. 2009. *Freebase*. San Francisco, CA: Metaweb Technologies. *http://www.freebase.com*. For data dumps, see *http://download.freebase.com/datadumps/*.

Garner, Ralph. 1963. "A computer-oriented graph theoretic analysis of citation index structures." In *Three Drexel Information Science Research Studies*, ed. Barbara Flood. Philadelphia, PA: Drexel Press.

González, Marta C., César A. Hidalgo, and Albert-László Barabási. 2008. "Understanding individual human mobility patterns." *Nature* 453: 779–782. doi: 10.1038/nature06958.

Henry, Nathalie, J-D. Fekete, and M. McGuffin. 2007. "NodeTrix: A hybrid visualization of social networks." *IEEE Transactions on Visualization and Computer Graphics* 13, no. 6: 1302–1309.

Henry, Nathalie. 2008. "Exploring large social networks with matrix-based representations." PhD diss., Cotutelle Université Paris-Sud and University of Sydney. *http://research.microsoft.com/en-us/um/people/nath/docs/Henry_thesis_oct08.pdf*.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Siemeon Warner. 2008. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. *http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm*.

Lancichinetti, A., and S. Fortunato. 2009. "Community detection algorithms: A comparative analysis." *Physical Review E* 80, no. 5, id. 056117. doi: 10.1103/PhysRevE.80.056117.

Leicht, E.A., and Raissa M. D'Souza. 2009. "Percolation on interacting networks." *arXiv* 0907.0894v1, *http://arxiv.org/abs/0907.0894v1*.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. "Network motifs: Simple building blocks of complex networks." *Science* 298, no. 5594: 824–827.

Nesselrath, Arnold. 1993. "Die Erstellung einer wissenschaftlichen Datenbank zum Nachleben der Antike: Der Census of Ancient Works of Art Known to the Renaissance." Habilitation thesis, Universität Mainz. Available at the CENSUS office at HU-Berlin.

Newman, Mark E.J. 2005. "Power laws, Pareto distributions and Zipf's law." *Contemporary Physics* 46: 323–351. doi:10.1080/00107510500052444.

Newman, Mark E.J., Albert-László Barabási, and Duncan J. Watts, eds. 2006. *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton University Press.

Penfield, W., and T. Rasmussen. 1950. *The Cerebral Cortex of Man: A Clinical Study of Localization of Function*. New York: Macmillan.

Phosphosite. 2003–2007. *PhosphoSitePlus™, A Protein Modification Resource*. Danvers, MA: Cell Signaling Technology. *http://www.phosphosite.org*.

Pietriga, Emmanuel, Christian Bizer, David Karger, and Ryan Lee. 2006. "Fresnel—A browser-independent presentation vocabulary for RDF." In *The Semantic Web—ISWC 2006*, vol. 4273, Chapter 12. Eds. I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo. Berlin, Heidelberg: Springer Berlin Heidelberg.

Saxl, Fritz. 1957. "Continuity and variation in the meaning of images." Lecture at Reading University, October 1947. In Lectures. London: Warburg Institute.

Schich, Maximilian. 2009. "Rezeption und Tradierung als komplexes Netzwerk. Der CENSUS und visuelle Dokumente zu den Thermen in Rom." Ph.D. diss., Humboldt-Universität zu Berlin. Munich: Verlag Biering & Brinkmann. urn:nbn:de:bsz:16-artdok-7002.

Schich, Maximilian, and Albert-László Barabási. 2009. "Human activity—from the Renaissance to the 21st century." In *Cultures of Change. Social Atoms and Electronic Lives. Exhibition Catalogue: Arts Santa Mònica, Barcelona, 11 December 2009 to 28 February 2010*. Gennaro Ascione, Cinta Massip, and Josep Perelló eds. Barcelona: Arts Santa Monica. urn:nbn:de:bsz:16-artdok-9582.

Schich, Maximilian, and Sybille Ebert-Schifferer. 2009. "Bildkonstruktionen bei Annibale Carracci und Caravaggio: Analyse von kunstwissenschaftlichen Datenbanken mit Hilfe skalierbarer Bildmatrizen." Project report. Rome: Bibliotheca Hertziana (Max-Planck-Institute for Art History). urn:nbn:de:bsz:16-artdok-7121.

Schich, Maximilian, César Hidalgo, Sune Lehmann, and Juyong Park. 2009. "The network of subject co-popularity in classical archaeology." urn:nbn:de:bsz:16-artdok-7151.

Schich, Maximilian, Sune Lehmann, and Juyong Park. 2008. "Dissecting the canon: Visual subject co-popularity networks in art research." 5th European Conference on Complex Systems, Jerusalem (online material). urn:nbn:de:bsz:16-artdok-7111.

Science. 2009. Special Issue on Complex Systems and Networks. *Science* 325, no. 5939: 357–504. *http://www.sciencemag.org/content/vol325/issue5939/#special-issue*.

Segaran, Toby. 2009. "Connecting data." In *Beautiful Data*. Sebastopol, CA: O'Reilly Media.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A software environment for integrated models of biomolecular interaction networks." *Genome Research* 13, no. 11: 2498–2504. doi: 10.1101/gr.1239303.
See also *http://www.cytoscape.org*.

Sullivan, Danny. 2005. "Search engine sizes." *Search Engine Watch*. *http://searchenginewatch.com/2156481*.

Wassermann, Stanley, and Katherine Faust. 1999. *Social Network Analysis: Methods and Applications, Fourth Edition*. Cambridge: Cambridge University Press.

Wikipedia. "Wikipedia: Size comparisons." *http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons*.

Winckelmann Corpus. 2000. *Corpus der antiken Denkmäler, die J.J. Winckelmann und seine Zeit kannten*. Winckelmann-Gesellschaft Stendal, ed. DVD and online database. Munich: Verlag Biering & Brinkmann/Stiftung Archäologie.