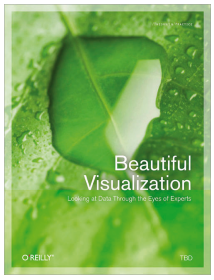


The present article will be published as a book chapter in
Beautiful Visualization. Looking at Data through the Eyes of Experts.
Edited by Julie Steele and Noah Iliinsky. Sabastopol, CA: O'Reilly. April 2010
ISBN: 978-1-4493-7986-5. URL: http://oreilly.com/catalog/0636920000617

Author contact: maximilian@schich.info

Revealing Matrices

Maximilian Schich

This chapter uncovers some non-intuitive 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 this is my trained area of expertise. However, what I present here—namely that the complex structure of databases can be visualized in a useful depiction—can also be shown for many other structured data collections, including biological research databases or 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 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 catalogues and bibliographies, image archives, museum inventories, and more general research databases. All of them can be built on extremely complicated data models, but even the most boring examples can be simple on the surface and, given enough data, confusingly complex in any single link relation. The thematic coverage potentially includes all man-made objects, such as in the Library of Congress Classification System, which 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 Ancient Monuments—such as Roman sculpture and architecture—appearing in Western Renaissance documents, such as sketchbooks, drawings, and guidebooks. In particular, we are looking at the last state of the database, before it was transferred from a graph-based database system (CENSUS 2005) to a more conservative relational database format (CENSUS BBAW) in 2006, allowing for comparison of the historic state with current and future achievements.

The more, the better?

After working in the field of 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 use of standards, user interfaces, fancy project titles, and recent buzz words in the project description. Regarding content, evaluators are often satisfied with a few basic measures such as the number of records in the database and 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 data quality collected and processed within their framework. The same is true for user interfaces, which give as much indication for the quality of content as does the aspect ratio of a paper sheet of text. Furthermore, data standards as well as 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 and even faster than a modern Python script. As a consequence, we should avoid any form of system patriotism in project evaluation, where the user of a particular standard has to be afraid of being evaluated by the fans of another.

Even standards, which we all consider desirable, such as Open Access are difficult to ponder: While Open Access provides a positive spin to many current projects, it is not entirely clear what it should mean within the realm of curated databases. Should we really be satisfied with a complicated but free user interface (cf. Bartsch 2008 fig. 10), or should we rather expect a sophisticated API and periodical dumps of the full database (cf. Freebase), which would allow for serious analysis and a 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?

Finally, we are forced to look at the actual content of any given project. We will find out within this chapter that it makes only limited sense to focus on the subtleties of a few particular entries within database evaluation, as usually there is no average amount of

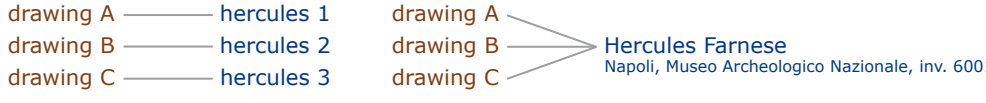
information related to any particular database entry. The omnipresent phenomenon of Long Tails (Anderson 2006, Newman 2005, Schich Hidalgo et al. 2009 note 5), which we will encounter in almost all figures below, hinders us to extrapolate from a few data-rich entries to the whole database—i.e. from the Pantheon in the CENSUS to all other Ancient Monuments.

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 their number of articles (cf. Wikipedia s.v. Size comparisons), Biomedical databases publish their number of compounds, genes, or proteins (cf. Phosphosite 2003-2007 or Flybase 2008), and even search engines traditionally (but ever more reluctantly) provide the number of pages in their index (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." (*http://www.census.de*, retrieved 9/14/2009)

Although these numbers are for sure impressive from the point of view of Art History, where large exhibition catalogues usually contain a couple of hundred entries, it is easy to disprove the significance of numbers of records as a really good measure of database quality, if taken in isolation. Just like 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 x-1 gives a striking example of this phenomenon, where the quality of a dataset grows while the number of entries in the dataset is reduced. 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 the example below.

FIGURE X-1. The quality of a dataset may grow, while reducing the number of records: A merge of three unspecified depictions of hercules (left) to a well identified statue (right) is very desireable, as it requires sufficient domain expertise in the curator and a well organized data processing pipeline.



In the end, we have to find more sophisticated measures in order to evaluate a given project. If we really want to know the value of a given dataset, we have to look at the global emerging structure, which is not foreseeable by looking at the commonly used indicators. The only thing we can expect in any dataset is that the global structure can be characterized as a non-trivial complex system. The complexity emerges from local activity (Chua 2005), as the availability and the attention towards the source data are highly hetereogeneous by nature and every curator has a different idea about the a priori datamodel definitions. As the resulting complexity is hard to predict, we have to measure and visualize it in a meaningful way.

Databases as Networks

Structured data in 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 or RDF graphs; semi-structured data is found in wikis, PDFs, HTML pages, and, more than elsewhere, on traditional paper. Disregarding the subtleties of all these representational forms, the underlying technical structure usually involves three areas:

1. a data model convention, ranging from simple index card separators in a wooden box to complicated ontologies in your favorite representational language;

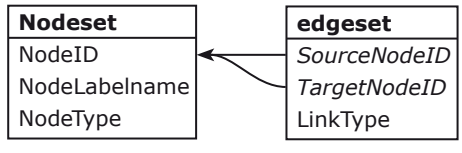
2. data formatting rules, including display templates such as lenses (Pietriga et al. 2006) or predefined query instructions; and

3. data processing rules that act according to these instructions.

Here, we are interested first and foremost how the chosen data model convention 1 interrelates with the available data.

As Toby Segaran (2009) pointed out in *Beautiful Data*, there are two ends of the spectrum regarding the data model convention. 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 x-2 that can support any type of data, essentially representing the data as a graph instead of a set of tables.

FIGURE X-2. Even the most complicated databases can be stored as a graph in a basic schema with two tables: Objects or records are stored as nodes in a nodeset table. Relations are stored as links or edges in an edgeset table. (cf. Figure 20-3 in Segaran 2009).



Represented in this form, every database can be considered a network. Database entries form the nodes of the network, database relations figure as the connections between the nodes, the so called edges or links. If we consider databases in art research as networks, the nodes are the entries representing physical objects such as Monuments and Documents, as well as Persons, Locations, Time ranges or Events (cf. Saxl 1947). As a consequence, a large number of possible node types emerges. A relation between two nodes—such as "Drawing A was created by Person B"—is a link, naturally forming 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, far beyond the state of knowledge at the point of time when the database was conceptualized, and unreachable by a regular local query. Consequently we can reach beyond the common measures of quality in evaluation. We can check how the data actually fits the data model convention; if the applied standards are appropriate; and if it makes sense to connect the database with other sources of data.

Data Model Definition Plus Emergence

In order to get an idea of the basic structure, the data model is the first thing we want to see in a database evaluation, 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 x-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 a RDF-dump (Freebase) or Linked Data (Bizer Heath 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 biologic network science community. The resulting diagram in Figure x-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 as shown in Figure x-3 is a meta network extracted from the graph database schema according to Figure x-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, 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 as well as 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 which is usually not reflected in traditional E-R data structure diagrams, often leading to lengthy discussions about almost irrelevant areas of particular database models.

The heterogeneity of node and link type frequency that we can see in Figure x-3 is not restricted to our example. It is observable in many datasets, as can be shown for research databases (Schich Ebert-Schifferer 2008), large bibliographies (Schich Hidalgo et al. 2009), Freebase or the Linked Data cloud, no matter whether the number of types is predefined or expandable by the curators. In all cases that I have seen so far, the number of nodes per nodetype as well as the number of links per linktype exhibit right-skewed diminishing distributions, which are widely known as *long tails* (Anderson 2006, Newman 2005), lacking a shared average as found in the 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 special issue 2009). Figure x-3 makes clear that the observed heterogeneity is also present at the level of node and link types within more structured data graphs.

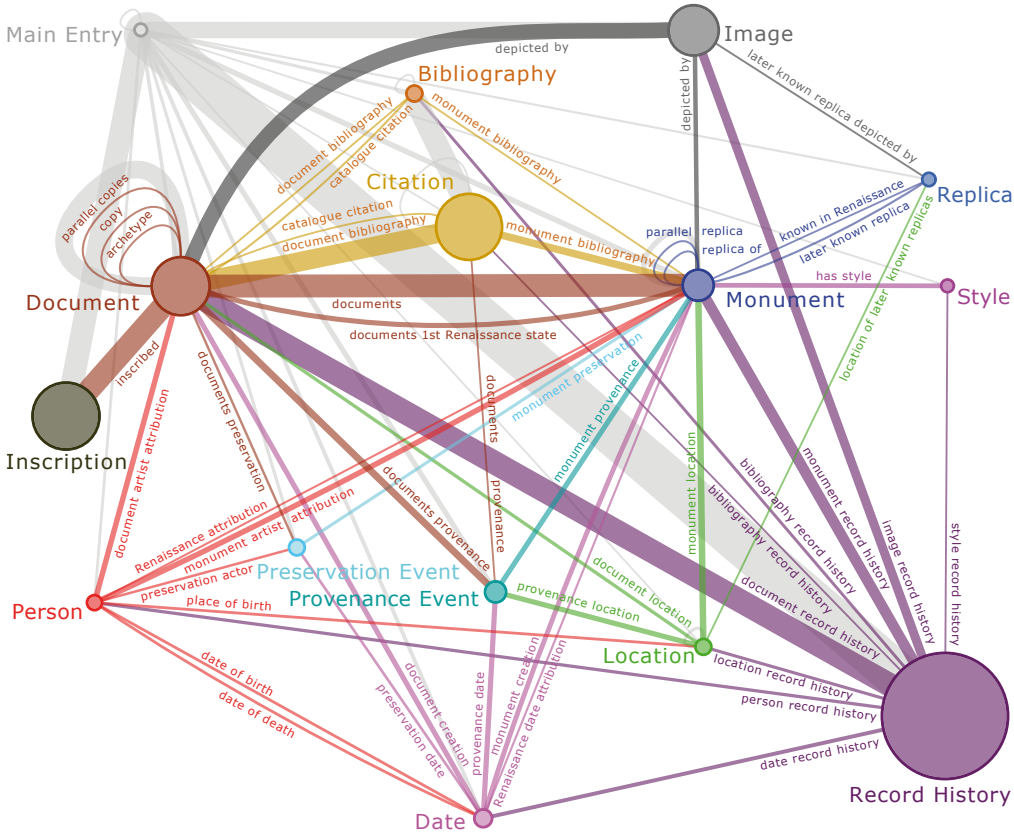


FIGURE X-3. The CENSUS data model depicted as a node-link diagram – equivalent to a regular Entity-Relationship (E-R) data structure diagram. The node size reflects the number of nodes per nodetype; the line width corresponds to the number of links per linktype. Both measures span several orders of magnitude.

Network Dimensionality

Looking more closely at Figure x-3, we 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, respectively differing by being the target or the source of the central documentation link. Whereas in general any physical object can function as a monument and as a document, the CENSUS divides them into discrete node types, due the fact that both groups belong to different periods, Classical Antiquity and Western Renaissance: Roman sculpture and architecture is documented by Renaissance drawings, sketchbooks, texts, etc.

In addition to these central dimensions, there is another node type representing physical objects called Replica, containing later-known replica Monuments, which 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

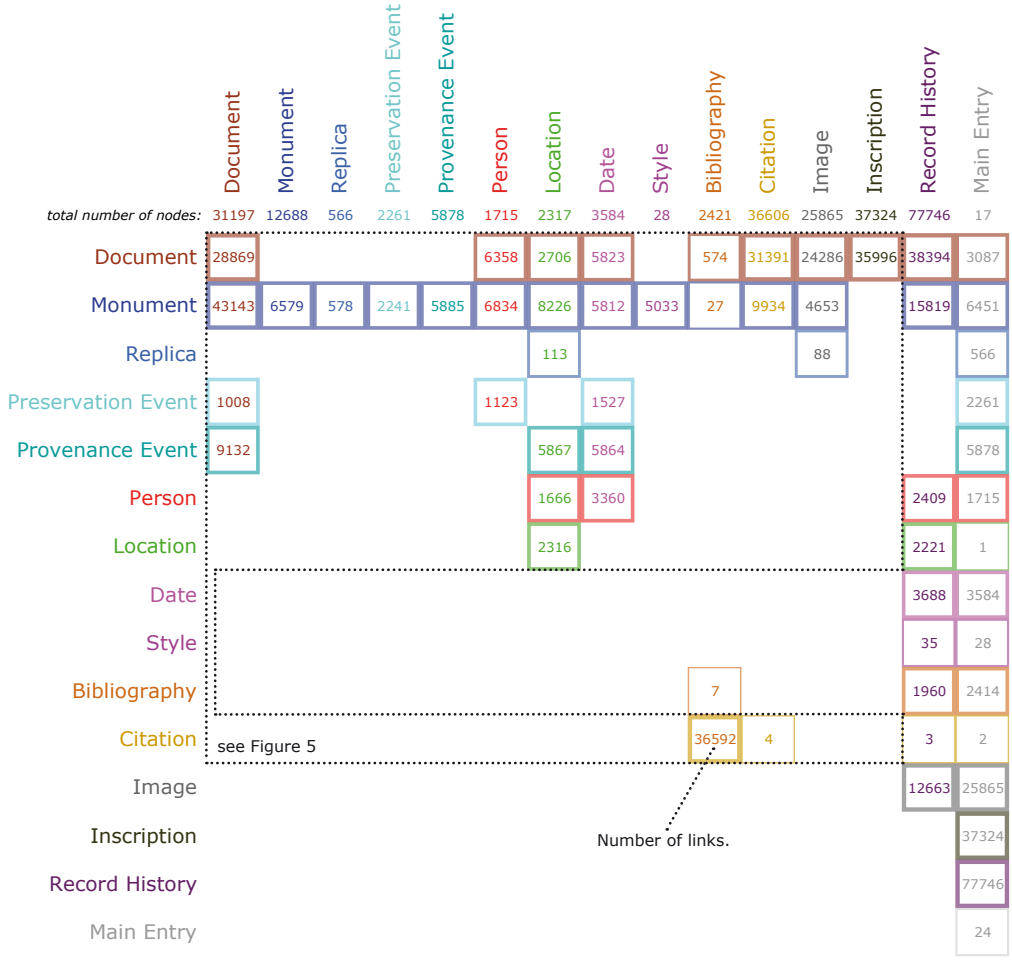


FIGURE X-4. The CENSUS data model depicted as an adjacency matrix. Every node type is represented as a line and a column. The total number of links between two nodetypes is indicated in the intersecting cell of the respective lines and columns. Links point OUT of the lines IN to the columns. The number of links given in the cells combines all existing link types between two nodetypes – for example 'date of birth' and 'date of death' in the Person–Date cell. The 'total number of nodes' given in the first row corresponds directly to the node size in Figure X-3.

all functions are defined by the presence of certain links pointing in or out of a particular node. Functionality constraints regarding relational databases, which caused the current setup, do not apply anymore as they did in the early 1980s, when the data model was initially conceived.

Distributed around the physical objects in Figure x-3 we find Persons, Locations and time ranges, such as Date and Style. Relations between all these dimensions are mostly modeled using direct links. Persons are 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.

A notable exception to these shortcuts are the Preservation and Provenance Events, stating 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 both 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 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 inherent as emergent star motifs, such as in the Linked Data graph.

The CENSUS becomes an authoritative—i.e. citable—source of information by providing a variety of meta dimensions, 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 within the re-conversion of the CENSUS to a relational database. In the former graph-based system, the Main Entries figured as database chapters due to the lack of tables in order to facilitate navigation by bundling all Persons, Locations, etc.

The Matrix Macroscope

The node-link diagram in Figure x-3 is only one possibility to depict the CENSUS data model. As any network consisting of nodes and links, we can also depict it in the form of a so-called *adjacency matrix* (cf. Garner 1967, Bertin 1981; Bertin 2001; Henry 2008), as shown in Figure x-4. Here, the node types are represented as the vertical columns and horizontal lines of a table, where link information appears in a cell across a particular line and column. Regarding the place of birth for example, you can imagine the link pointing out of the Person line 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 x-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 x-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, such as "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 x-5, for example, we see a node-link diagram of all the nodes and links occurring between two node types in a cell. We do so by using a layout algorithm, which is relatively inexpensive from a computational point of view, such as the y-files organic layout, which is a part of the Cytoscape application. As a consequence, all the explicit node-link data in the whole database appears in the data model matrix.

Looking at the result in Figure x-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 dyads connecting two nodes exclusively. Another thing we 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 in 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, which is rooted at 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. However, its prominence reminds me of the oversized hands in the senso-motory *homunculus* of the human brain (Penfield Rasmussen 1950; Dawkins 2003 p. 243-244)—the CENSUS seems to have a *romunculus*. In the same way as an overly large area of our brain's cortex is dedicated to hand coordination and the sense of touch in our hands, the CENSUS location hierarchy seem to be biased towards sculpture collections in Rome. Like a master pianist, whose hands occupy even more space on the cortex than in regular people, the CENSUS seems to be formed by specialization – such as by the addition of Ulisse Aldroandi's famous books (1556 and 1562) that list thousands of sculptures in Roman collections (cf. Schich 2009 p. 124/125).

Another interesting feature of Figure x-5 are the disproportionally large stars, which we can find in a number of cells. Some of the stars are natural properties of the data, as in the case of the 11927 Document nodes linked to the Bibliographic node Bartsch 1854-

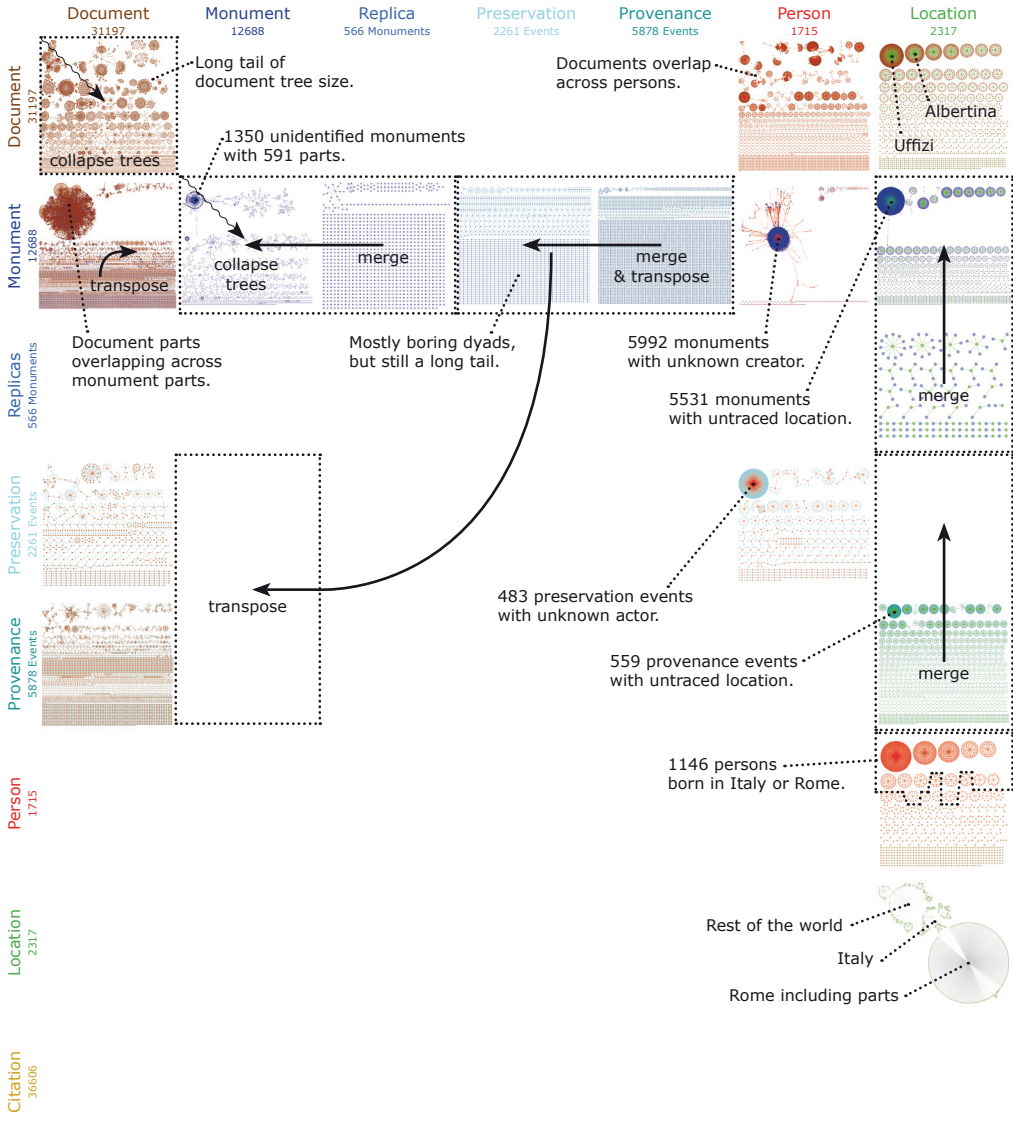
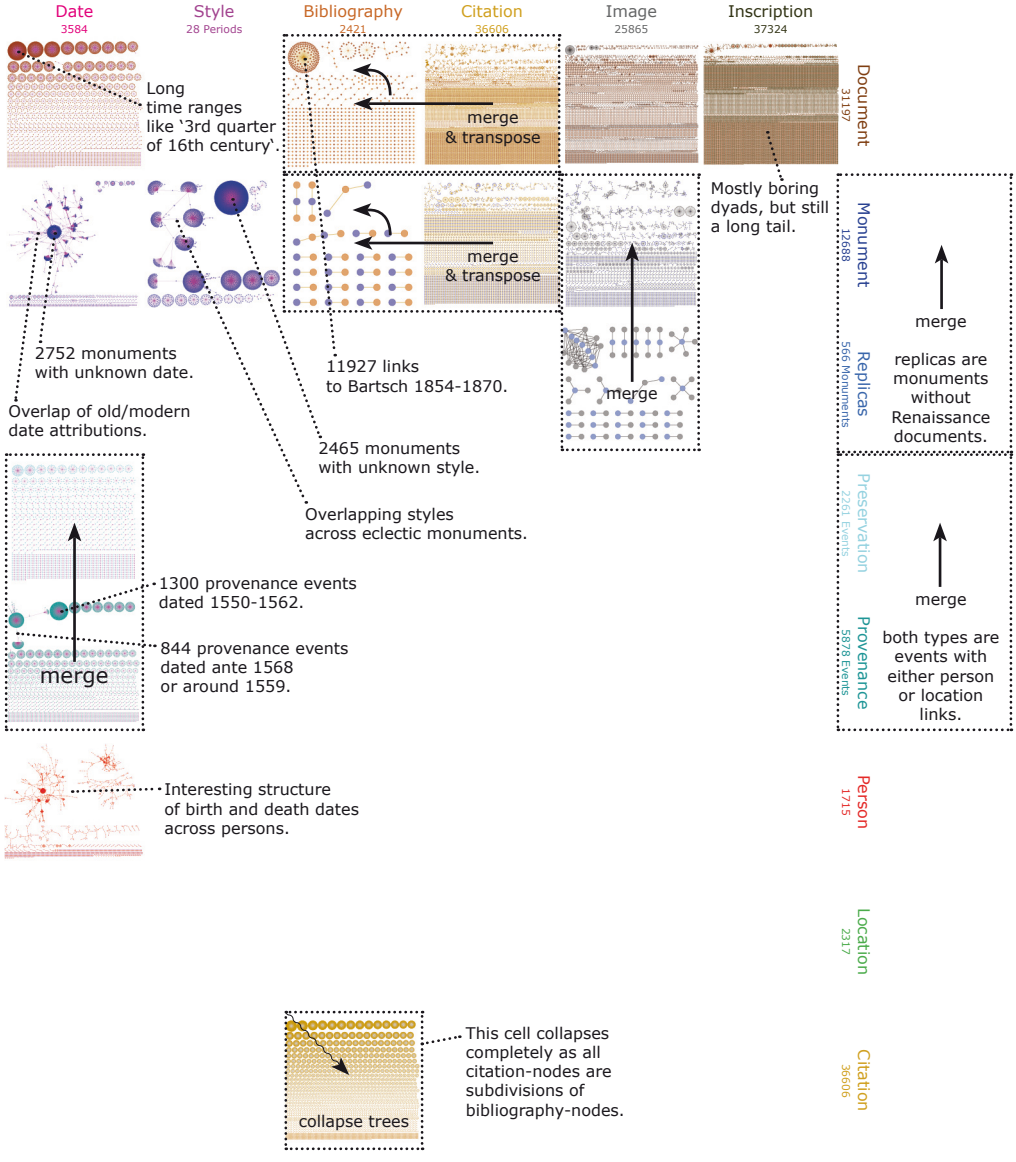


FIGURE X-5. An enriched subsection of the CENSUS data model, as marked Figure X-4. A node-link diagram of all relevant links is now depicted in every cell. Emerging data structure becomes visible within the data model definition. In most cases we can see a long tail of shrinking star motifs that fade out to a large number of simple dyads. Only in some cases, such as the Monument–Document cell, the visible structure appears to be more complex.

1870, or as in the case of the 1146 Persons born in Italy or Rome. However, most of the giant stars are artifacts related to unknown entries, such as 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 1350 unidentified Monuments, 5992 Monuments with unknown creator,



5531 monuments with untraced location, 2752 Monuments with unknown Date, 2465 Monuments with unknown Style, 483 Preservation Events with unknown actor, and 559 Provenance Events with untraced Location. To be sure, the presence of all these unknown entries is not an error; the attribution of an unknown Date could, for example, refute a wrong Renaissance date attribution. However, the numbers provide a feeling for how incomplete our knowledge is. Also, 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 to Figure x-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. So 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 rarely communicated about this database before. The same is true for Monuments. Here again, a small percentage of the records, in particular Architecture, is subdivided into trees including building parts, rooms, and even tiny individual features of architectural decoration. A third area including equivalent subdivisions in the database is the Bibliography, which is effectively subdivided into Citations, such as paragraphs of text in modern scholarly books.

The consequence of all these subdivisions in Figure x-5 is that particular links point from and to particular sub-nodes: from Monument parts to Document parts instead from entire Monuments to the entire Documents; 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 in how many sketchbooks a group of Monuments appears (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 x-6 and redraw the entire matrix as in Figure x-7(a).

Collapsing the trees of Documents, Monuments and Bibliographic Citation to single nodes works in the following way (cf. Schich 2009 p. 28-37): In Figure x-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 x-6(a‘). In order 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, the more subnodes it contains in its collapsed tree; the broader a line 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 x-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 6(c)/(c‘).

The most striking feature of the refined cell in Figure x-6(c‘) is the emergence of a so-called Giant Connected Component (GCC), which connects almost 90% of all Monu-

ments 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 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 guide books, sketchbooks and city maps. A surprising feature in the periphery of the GCC is the dominance of brush-like structures connected to large Document nodes: Obviously a major fraction of all Monuments in the CENSUS is connected to only one single Document, either because the Documents lack sufficient information or because they were not identified and normalized by the curators for other reasons.

As 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 x-5: As with any adjacency matrix, we can sort (or *permutate*) the columns and lines along the horizontal and vertical axis, without loosing any information (Bertin 1981; Bertin 2001). We can 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 x-7(a)-(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 citation, 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 true for the number of Events per Monument, of

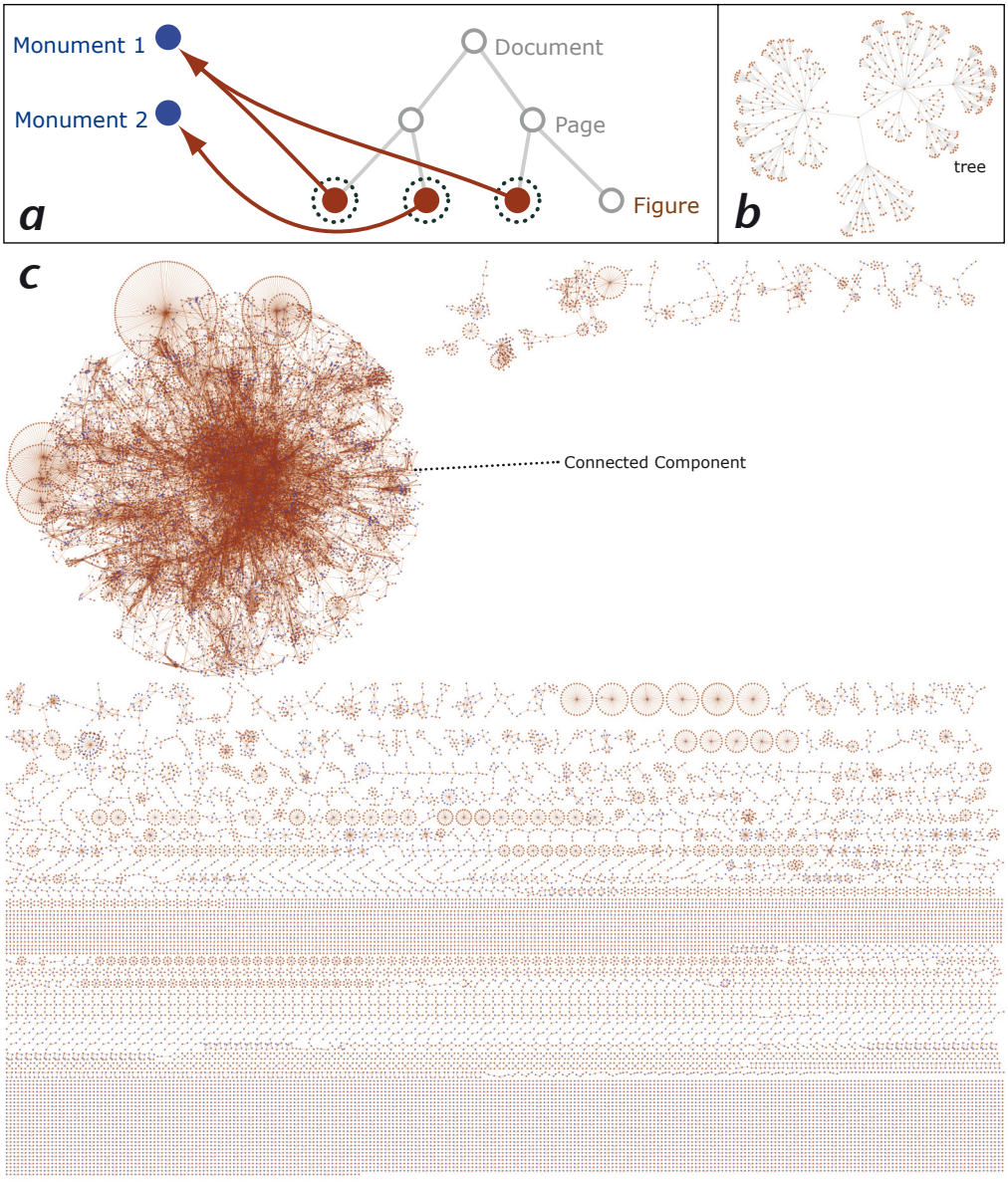


FIGURE X-6. Looking at the raw CENSUS data, Documents, Monuments and Bibliographic Citations are too fragmented to draw meaningful quantitative conclusions. a) Single parts of larger Documents are connected to single Monument nodes; b) Single parts of larger Documents are held together by 'parent links' forming a forest of Document trees in the Document–Document cell (cf. Figure X-5); c) In the raw Monument–Document cell the single parts of larger Documents are separate. Surprisingly there is nevertheless a fairly large connected component of overlapping document parts, i.e. text paragraphs and overview figures that are connected to multiple Monuments – forming a small hairball. However, the structure of the connected component looks rather random, without many interesting features.

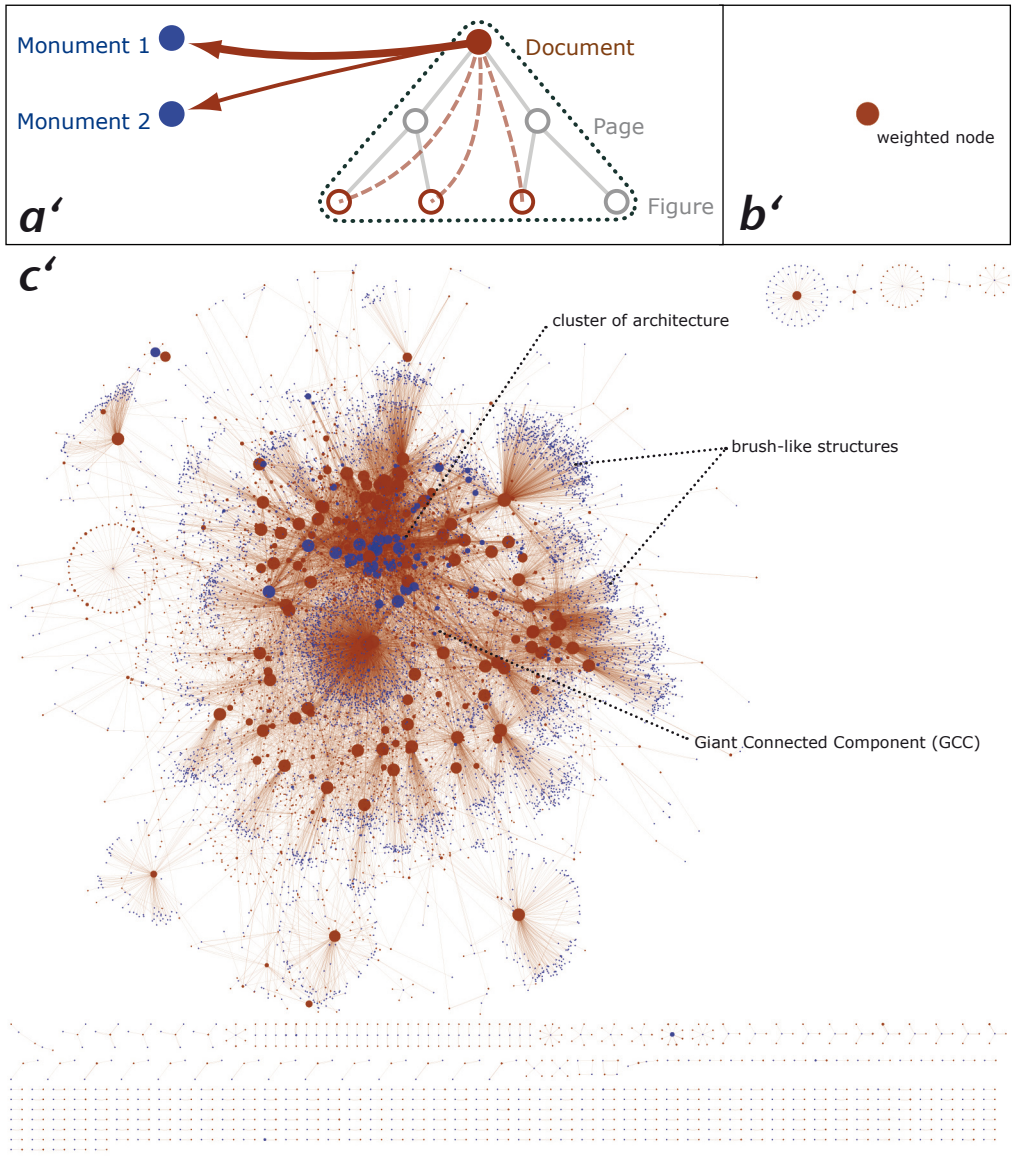


FIGURE X-6. (continued) If we reduce trees of large Documents, Monuments, and bibliographic Citations to single nodes, a transformation takes place, and more sophisticated complex structure in the database becomes apparent. a‘) After the tree collapse, the entire Document is now represented as a single node, connected to the monuments via weighted links. The link weight corresponds to the collapsed number of links between two collapsed nodes; b‘) The collapsed Document, Monument and Bibliographic Citation trees are now represented as weighted nodes; c‘) In the reduced Monument–Document cell, 89.9% of all collapsed nodes are connected in a single ‘Giant Connected Component (GCC)‘. Brush like structures in the diagram indicate a very large number of Monuments, characterized by a lack of sufficient information, subsequently connected to only a single Document. Architectural Monuments, easily recognizable by their large node weight due to their subdivisions, cluster together as their coverage overlaps across documents, for example across guide books and city maps.

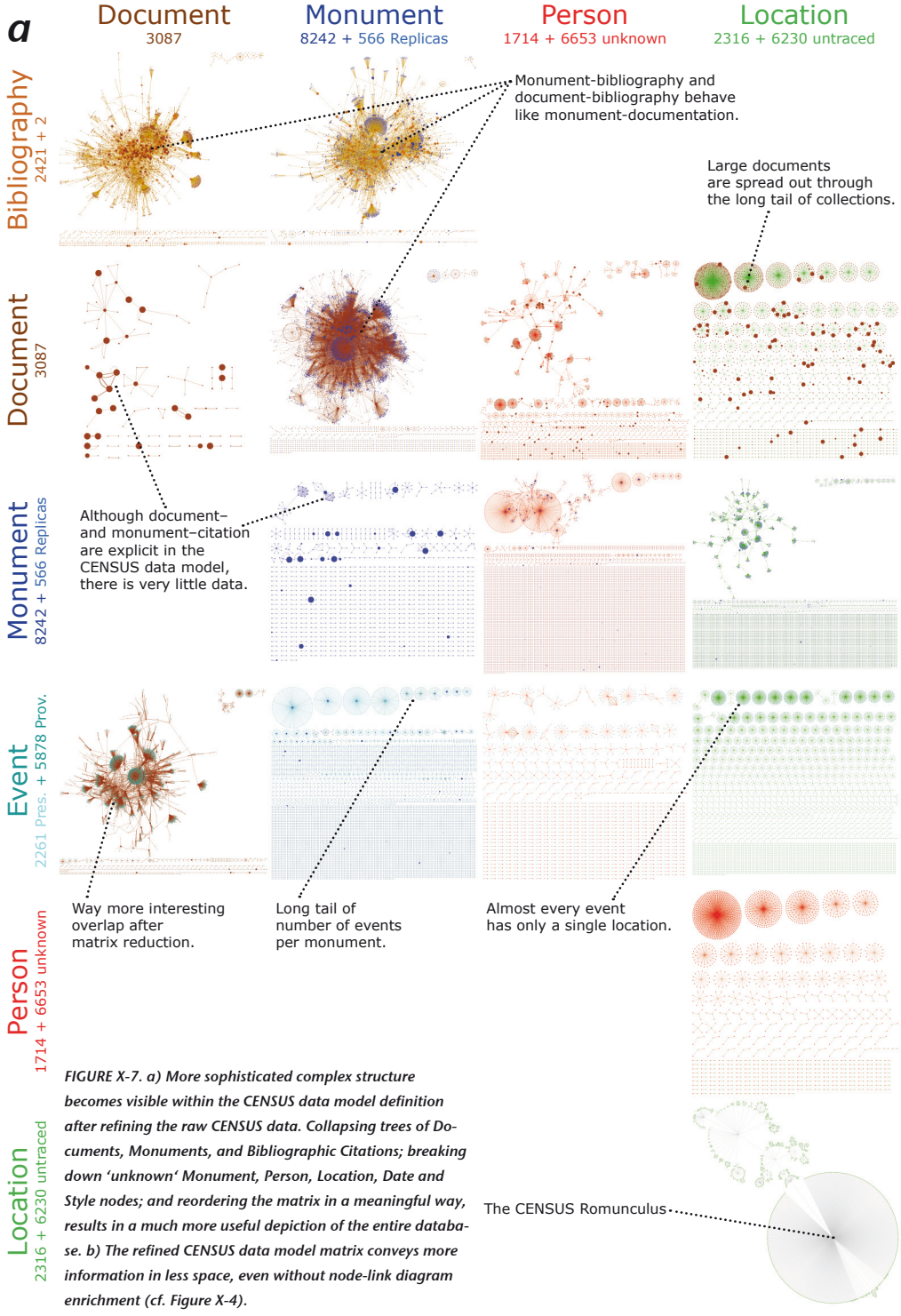
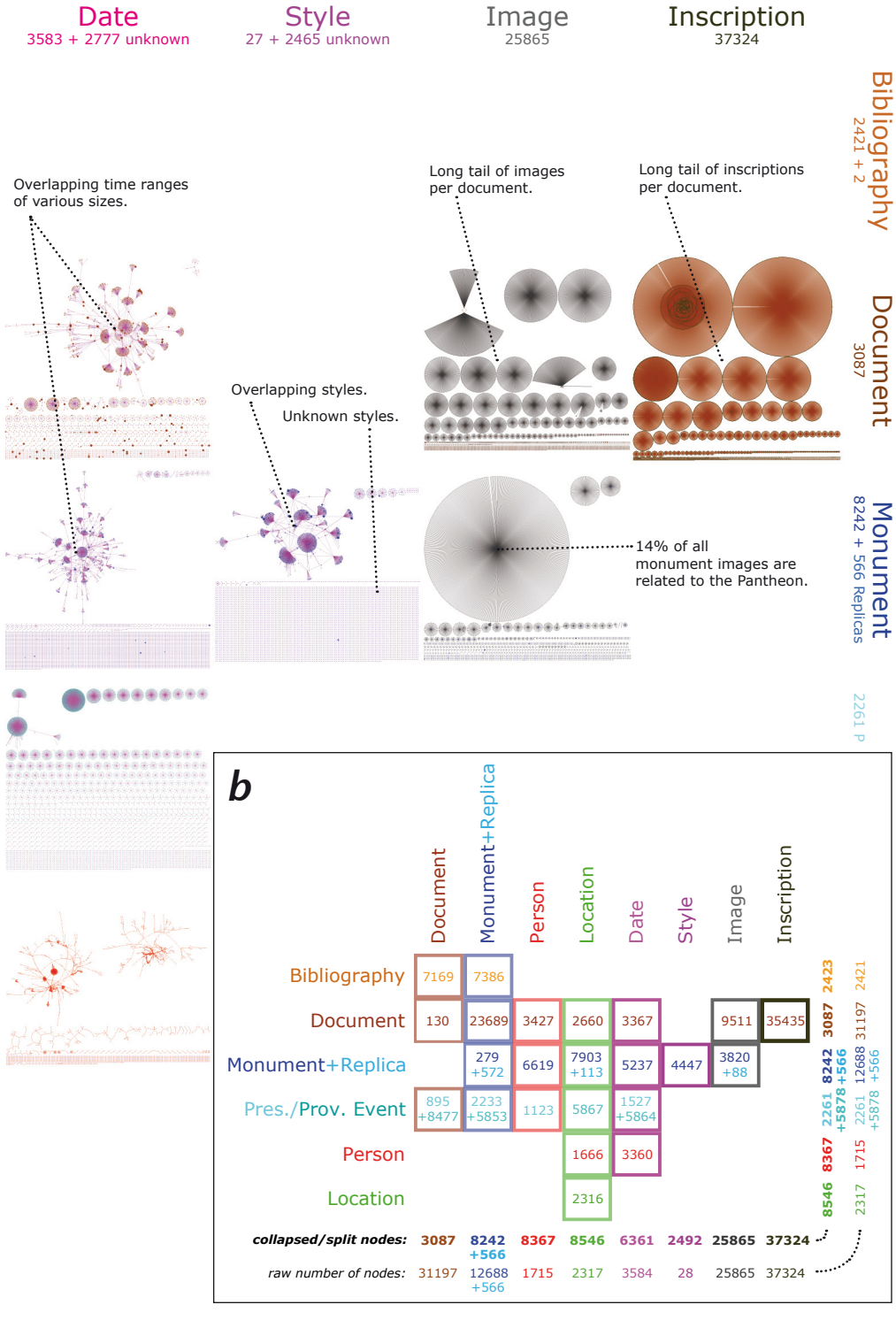


FIGURE X-7. a) More sophisticated complex structure becomes visible within the CENSUS data model definition after refining the raw CENSUS data. Collapsing trees of Documents, Monuments, and Bibliographic Citations; breaking down 'unknown' Monument, Person, Location, Date and Style nodes; and reordering the matrix in a meaningful way, results in a much more useful depiction of the entire database. b) The refined CENSUS data model matrix conveys more information in less space, even without node-link diagram enrichment (cf. Figure X-4).


Date
3583 + 2777 unknown
Style
27 + 2465 unknown
Image
25865
Inscription
37324
Bibliography
2421 + 2
Document
3087
Monument
8242 + 566 Replicas
2261 P
Overlapping time ranges of various sizes.
Overlapping styles.
Unknown styles.
Long tail of images per document.
Long tail of inscriptions per document.
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
collapsed/split nodes:
raw number of nodes:
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
3087
8242 +566
8367
8546
6361
2492
25865
37324
31197
12688 +566
1715
2317
3584
28
25865
37324
2423
2421
3087
31197
8242 +566
12688 +566
2261 +5878
2261 +5878
8367
1715
8546
2317

Images per Document/Monument, of Inscriptions per Document, or Events per Location. An interesting case regards the Document-Location cell, where we can see that large Documents are spread out 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 in 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 contains the most complex overlap, as it 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 x-7(a), is not feasible if the data becomes an order of magnitude larger than the CENSUS or even encompasses 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 x-8, we plot a cumulative IN- and OUT-degree-distribution (Broder et al. 2000; Newman 2005) for every linktype occurring in a matrix cell. As every link points OUT of the source node type 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 3000 on the x-axis. (In a regular linear projection, the slope of each distribution would be so steep that we could 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 size of people. Instead, we find a whole zoology of long tails ranging from beautiful power-laws to log-linear curves mixed with less clean, more bumpy 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, 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.

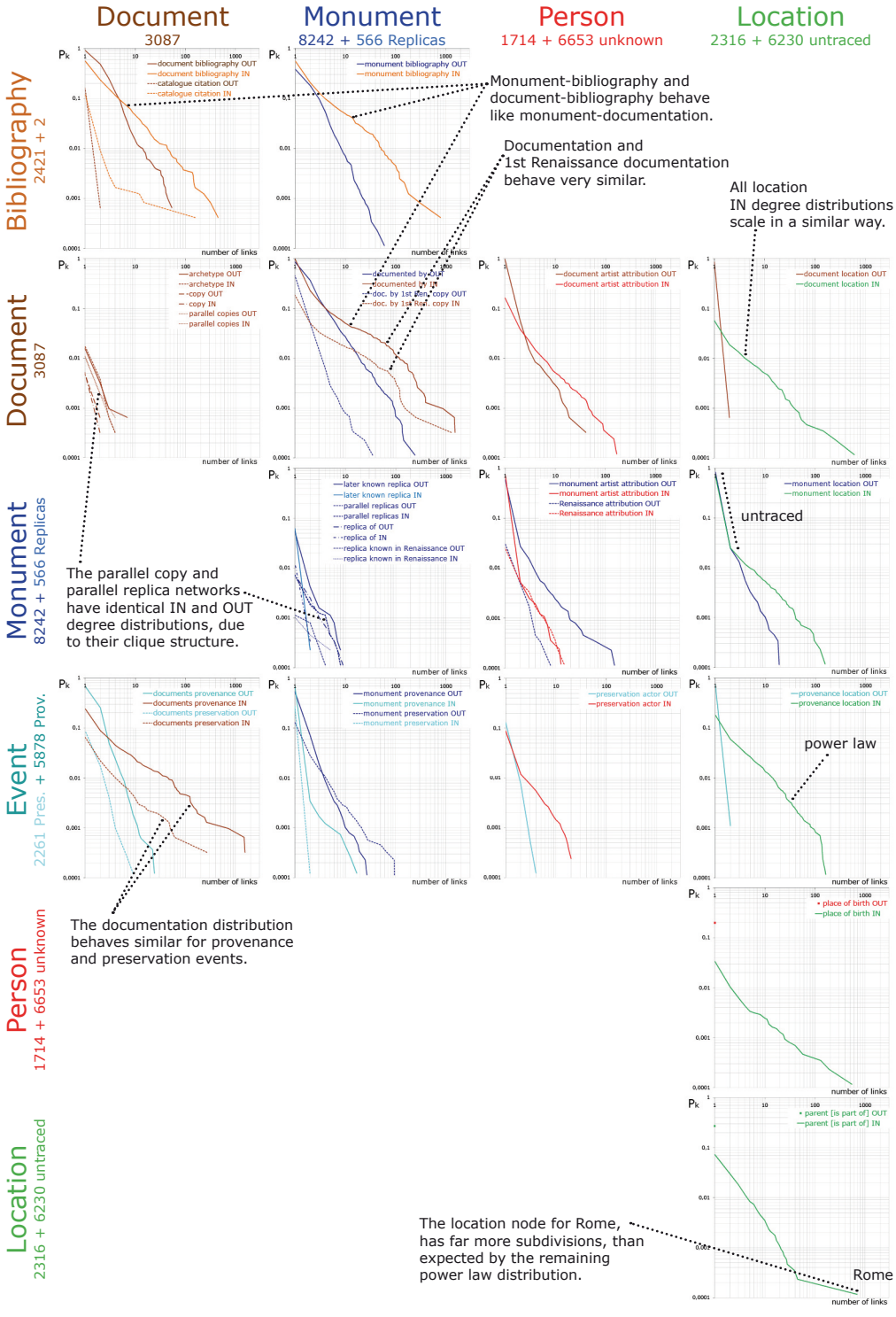
The only symmetric link relationship in the CENSUS can be found in the parallel copy and parallel replica links in the Document-Document and Monument-Monument cells respectively. Ideally, the IN and OUT degree distributions should be identical, as the relevant nodes are fully connected to so called cliques. In reality, both link types become more asymmetric the further down we go in 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 connect n parallel copies with n times (n-1) links among each other in a manual way.

Similarly, the behavior of certain relationships that we spotted in Figure x-7, such as the equivalence of Monument-Bibliography and Monument-Documentation, is confirmed in Figure x-8 (cf. Schich Barabási 2009). Not only is there an obvious similarity between these cells, but also 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 as well as preservation documentation in the Event-Documentation cell. The Locations IN degree scales in a very similar way across all relevant cells. An exception to this observed regularity is 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. All we have to do is look at the value where the respective curve crosses the y-axis. Accordingly, we can see 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 which were scanned by the CENSUS publishing partner in 1994 were still not linked in the database by 2005.

Further Applications

The presented visualizations serve as a starting point for a variety of activities: Besides the evaluation of particular project goals by funders and project leaders, further study includes the identification of interesting research topics: Every single cell in the matrix could be the subject of an extensive investigation, as exemplified in my Ph.D., 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 2008), or to study the effects of network interaction (Leicht 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 in case of the CENSUS, or databases that can be mapped to the same standard, such as the CIDOC CRM.


Document
3087
Monument
8242 + 566 Replicas
Person
1714 + 6653 unknown
Location
2316 + 6230 untraced
Bibliography
2421 + 2
Document
3087
Monument
8242 + 566 Replicas
Event
2261 Pres. + 5878 Prov.
Person
1714 + 6653 unknown
Location
2316 + 6230 untraced
Monument-bibliography and document-bibliography behave like monument-documentation.
Documentation and 1st Renaissance documentation behave very similar.
All location IN degree distributions scale in a similar way.
The parallel copy and parallel replica networks have identical IN and OUT degree distributions, due to their clique structure.
untraced
power law
The documentation distribution behaves similar for provenance and preservation events.
The location node for Rome, has far more subdivisions, than expected by the remaining power law distribution.
Rome
number of links

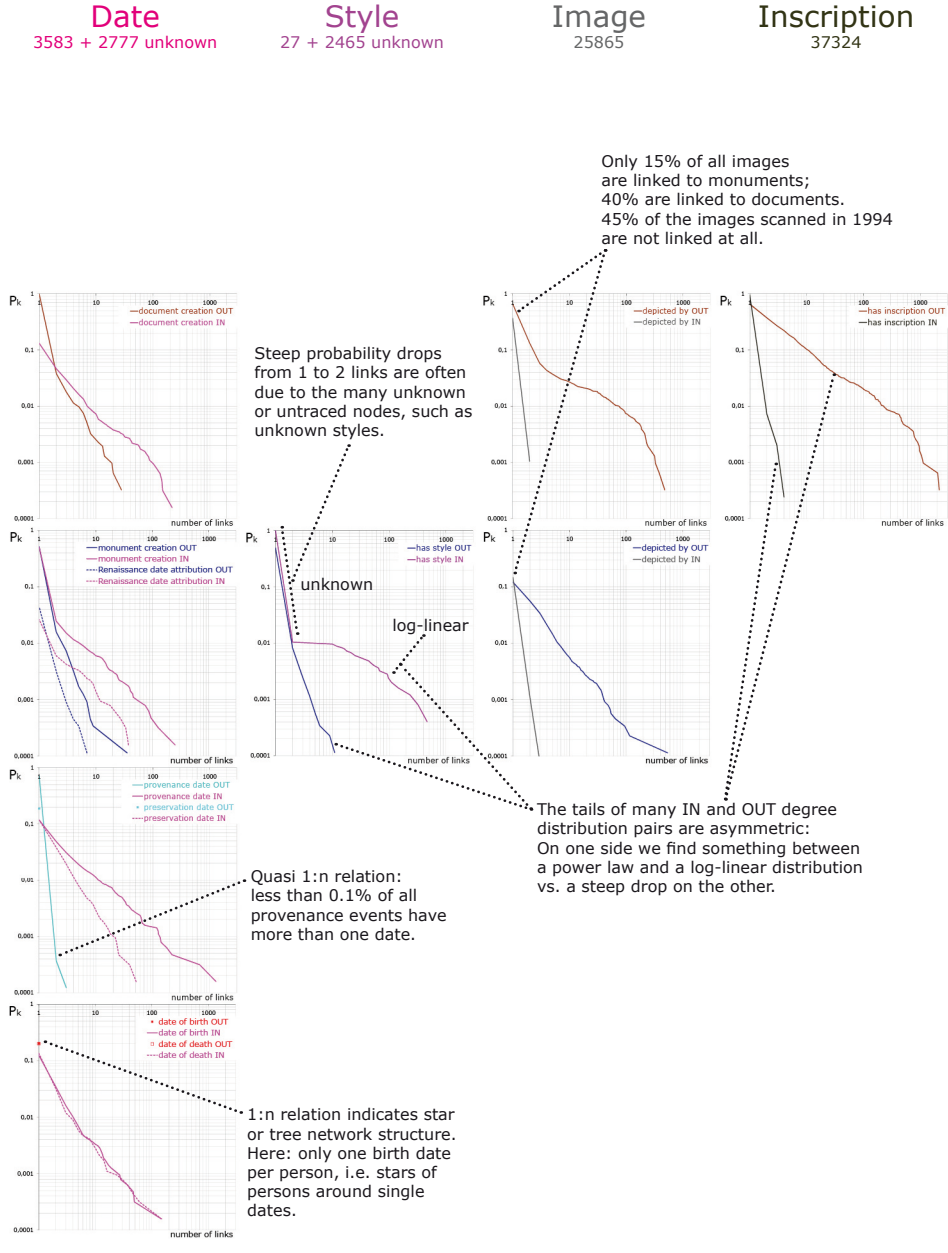


FIGURE X-8. The refined CENSUS data model matrix enriched with cumulative IN and OUT degree distribution plots, for every link type between two node types. Every plot indicates the cumulative probability P(k) (between 1 and 0.0001 on the y-axis) that a node has a certain number of OUT-going or IN-coming links (between 1 and 3000 on the x-axis). The axes are scaled logarithmic. The depiction reveals a whole zoology of tailed distributions. Whereas node-link diagrams as shown in Figures X-5 and X-7 are of limited use for larger datasets, a data model matrix using degree distributions or other network measures is easily scalable and useful regarding massive datasets.

Instead of dissecting a database in the way we just discussed here, it may also be interesting to combine separate networks in a similar visualization. Candidates for such a combination can be found in the social sciences or in biology, where networks such as citation, co-authorship, and image-tagging, or gene-transcription, protein-protein interaction, gene-disease, and others are only parts of a multipartite universe of conceivable networks.

The coarse graining of networks we looked at by collapsing the inherent trees of Documents, Monuments, and Bibliography, can also be done in many other different ways, for example by concentrating on particular subtrees or with more sophisticated methods such as blockmodelling (cf. Wassermann Faust 1994 pp. 394-424) or community finding (cf. Lancichinetti Fortunato 2009; Ahn Lehmann Bagrow 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 2007) or scalable image matrices (Schich et al. 2008) in relevant cells of the data model matrix.

Conclusion

In this chapter we found out that the use of enriched and refined data model matrices is very useful for database project evaluation, exposing many non-intuitive 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 within project evaluation by funders and the projects themselves, within a very short timeframe in a mostly automated process.

In our particular example of the CENSUS database, our visualization presents the first comprehensive big picture of the whole 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 the numbers given in the project description are 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 useful feedback I like to thank my audiences at NetSci‘09 in Venice and SciFoo09 in Mountain View, as well as my colleagues at the BarabásiLab at Northeastern University in Boston. Further thanks go to Stiftung Archäologie in Munich for providing the data and 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; Bagrow, James P.; Lehmann, Sune. 2009. "Link communities reveal multi-scale complexity in networks." *arXiv* 0903.3178 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. Mauro, Lucio. Venice.

Anderson, P. W. 1972. "More Is Different." *Science* v. 177 no. 4047 pp. 393-396.

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

Bartsch, Adam von. 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* v. 10 pp. 223–260

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

Bertin, Jacques. 2001. "Matrix theory of graphics." *Information Design Journal* v. 10.1 pp. 5-19. doi: 10.1075/idj.10.1.04ber

Bizer, Christian; Heath, Tom; Berners-Lee, Tim. 2009. "Linked Data – The Story So Far." *International Journal on Semantic Web & Information Systems.* v. 5.3 pp. 1-22. Preprint: http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

Broder, Andrei; Kumar, Ravi; Maghoul, Farzin; Raghavan, Prabhakar; Rajagopalan, Sridhar; Stata, Raymie; Tomkins, Andrew; Wiener, Janet. 2000. "Graph structure in the web." *Computer Networks* v. 33,1-6 pp. 309-319 doi:10.1016/j.physletb.2003.10.071

Butts, Carter. 2009. "Revisiting the Foundations of Network Analysis." *Science* v. 325.5939 pp. 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: Morgan Kaufmann.

Chen, Peter P.S. 1976. "The entity-relationship model—toward a unified view of data," *ACM Transactions on Database Systems (TODS).* v. 1.1 1-36. doi: 10.1145/320434.320440

Chua, Leon O. 2005. "Local Activity is the Origin of Complexity." *International Journal of Bifurcation and Chaos.* v. 15 pp. 3435-3456. doi: 10.1142/S0218127405014337

CIDOC-CRM. 2006. Crofts, Nick; Doerr, Martin; Gill, Tony; Stead, Stephen; Stiff, Matthew (eds.). *Definition of the CIDOC Conceptual Reference Model. 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. Auer, Sören; Bizer, Christian; Idehen, Kingsley (admins.): *DBpedia.* Leipzig: Universität Leipizig, Berlin: Freie Universität Berlin, Burlington/MA: OpenLink Software. http://dbpedia.org

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

Flybase. 2008. Drysdale, Rachel and the FlyBase Consortium. "FlyBase." *Drosophila.* pp. 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: Metaweb Technologies. http://www.freebase.com. Data dumps see http://download.freebase.com/datadumps/

Garner, Ralph. 1963. "A Computer-Oriented Graph Theoretic Analysis of Citation Index Structures." *Three Drexel Information Science Research Studies.* Philadelphia: Drexel Press. pp. 3-36

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

Henry, Nathalie; Fekete, J-D.; McGuffin, M. 2007. "NodeTrix: a Hybrid Visualization of Social Networks." *IEEE Transactions on Visualization and Computer Graphics (Proceedings of Visualization/Information Visualization 2007)* v. 13(6) pp. 1302-1309.

Henry, Nathalie. 2008. *Exploring Large Social Networks with Matrix-Based Representations.* Ph.D. Thesis, Cotutelle Université Paris-Sud and University of Sydney. http://research.microsoft.com/en-us/um/people/nath/docs/Henry_thesis_oct08.pdf

Lancichinetti, A.; Fortunato S. 2009. "Community detection algorithms: a comparative analysis" *Physical Review E.* v. 80, 056117. doi: 10.1103/PhysRevE.80.056117

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

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; Alon, U. 2002. "Network Motifs: Simple Building Blocks of Complex Networks." *Science* v. 298 pp. 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* v. 46 pp. 323-351. doi:10.1080/00107510500052444

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

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

Penfield, W. and Rasmussen T. 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; Bizer, Christian; Karger, David; Lee, Ryan. 2006. "Fresnel – A Browser-Independent Presentation Vocabulary for RDF." *5th International Semantic Web Conference.* Athens/GA.

Saxl, Fritz. 1947. "Continuity and Variation in the Meaning of Images." *Lectures.* London: Warburg Institute. printed 1957.

Schich, Maximilian; Lehmann, Sune; Park, Juyong. 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

Schich, Maximilian. 2009. *Rezeption und Tradierung als Komplexes Netzwerk. Der CENSUS und visuelle Dokumente zu den Thermen in Rom.* (Ph.D. Thesis, Humboldt-Universität zu Berlin, May 2, 2007) Munich: Verlag Biering & Brinkmann 2009. urn:nbn:de:bsz:16-artdok-7002

Schich, Maximilian; Ebert-Schifferer, Sybille. 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; Hidalgo, César; Lehmann, Sune and Park, Juyong. 2009. "The Network of Subject Co-Popularity in Classical Archaeology." To appear in the first issue of *Bolletino di Archaeologia On-line.* Preprint: urn:nbn:de:bsz:16-artdok-7151

Schich, Maximilian; Barabási, Albert-László. 2009. "Human activity – from the Renaissance to the 21st century." *Cultures of Change.* Exhibition Catalogue. ed. Perelló, Josep; Alsina, Pau. Barcelona: Arts Santa Monica 2009-2010.

Science special issue. 2009. "Complex Systems and Networks" *Science.* v. 325.5939 pp. 357-504. http://www.sciencemag.org/content/vol325/issue5939/#special-issue

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

Shannon, Paul; Markiel, Andrew; Ozier Owen; Baliga, Nitin S.; Wang, Jonathan T.;Ramage Daniel; Amin, Nada; Schwikowski, Benno; Ideker, Trey. 2003. „Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome Research.* v. 13.11 pp. 2498–504. doi: 10.1101/gr.1239303 – see also http://www.cytoscape.org

Sullivan, Danny. 28 January 2005. "Search Engine Size." *Search Engine Watch.* London: Incisive Media Ltd. http://searchenginewatch.com/2156481

Wassermann, Stanley; Faust, Katherine. 1994. *Social Network Analysis. Methods and Applications.* Cambridge: Cambridge University Press. (fourth edition 1999).

Wikipedia s.v. Size_comparisons. January 3, 2010. *Wikipedia. The free encyclopedia. English version.* http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

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

Chapter summary

Chapter X, *Revealing Matrices*, by Maximilian Schich, uncovers non-intuitive complex structures in curated databases, using data model definition matrices, enriched and refined with actual data.

Contributor

Maximilian Schich is an art historian working as DFG visiting research scientist at BarabásiLab, Center for Complex Network Research at Northeastern University in Boston, where he collaborates with network scientists, studying complex networks in art history and archaeology. Maximilian obtained his Ph.D. in 2007, and looks back at over a decade of consulting experience working with network data in art research, brokering within the tetrahedron of project-partners, users, programmers and customers. He worked several years with Projekt Dyabola as well as within Bibliotheca Hertziana (Max-Planck Institute for Art History), the Munich Glyptothek, and Zentralinstitut für Kunstgeschichte. You find more of his work at *http://www.schich.info*.

Author contact

Dr. Maximilian Schich
CCNR, Northeastern University
110 Forsyth St., 111 Dana
Boston, MA 02115
email: maximilian.schich.info
tel.: +1 (617) 817-7880