Big Data—Conceptual Modeling to the Rescue (Extended Abstract)

David W. Embley[1] and Stephen W. Liddle[2]

[1] Department of Computer Science
[2] Information Systems Department
Brigham Young University, Provo, Utah 84602, USA
embley@cs.byu.edu, liddle@byu.edu

1 Big Data

Every day humans generate several petabytes of data [ZEd$^+$11] from a variety of sources such as orbital weather satellites, ground-based sensor networks, mobile computing devices, digital cameras, and retail point-of-sale registers. Companies, governments, and individuals store this data in a wide variety of structured, semistructured, and unstructured formats. However, most of this data either languishes in underutilized storage repositories or is never stored in the first place. Ironically, in an era of unprecedented access to a veritable gold mine of information, it is increasingly difficult to unlock the value stored within our data. The essential problem of "Big Data" is that we are accumulating data faster than we can process it, and this trend is accelerating.

The so-called "four V's" characterize Big Data:

- *Volume*: applications sometimes exceeding petabytes[3]
- *Variety*: widely varying heterogeneous information sources and hugely diverse application needs
- *Velocity*: phenomenal rate of data acquisition, real-time streaming data, and variable time-value of data
- *Veracity*: trustworthiness and uncertainty, beyond the limits of humans to check

We should expect conceptual modeling to provide some answers since its historical perspective has always been about structuring information—making its *volume* searchable, harnessing its *variety* uniformly, mitigating its *velocity* with automation, and checking its *veracity* with application constraints. We do not envision any silver bullets that will slay the "werewolf" of Big Data, but conceptual modeling can help, as we illustrate with an example from our project that seeks to superimpose a web of knowledge over a rapidly growing heterogeneous collection of historical documents whose storage requirements are likely to eventually exceed many exabytes.

[3] Having successfully communicated the terms "mega-," "giga-," and "tera-byte," in the Big Data era we now need to teach users about "peta-," "exa-," "zetta-," and even "yotta-bytes." The NSA data center being built in Utah within 35km of our university purportedly is designed to store at least zettabytes (10^{21} bytes) and perhaps yottabytes (10^{24} bytes) of data.

2 Conceptual Modeling to the Rescue

Can conceptual modeling "come to the rescue" in some sense and help address some of the challenges of Big Data? We believe that the answer is affirmative. We do not expect conceptual modeling to address such issues as how we physically store and process bits of data, but "Moore's Law"[4] gives us confidence that future hardware technology will also help address these challenges. The mapping of conceptual models to efficient storage structures, a traditional application of conceptual modeling, is likely to be vastly different for Big Data and may be of some interest. But logical-to-physical design is not where we see the impact of conceptual modeling on Big Data. We expect that conceptual modeling can help by conceptual-model-based extraction for handling *volume* and *velocity* with automation, by inter-conceptual-model transformations for mitigating *variety*, and by conceptualized constraint checking for increasing *veracity*.

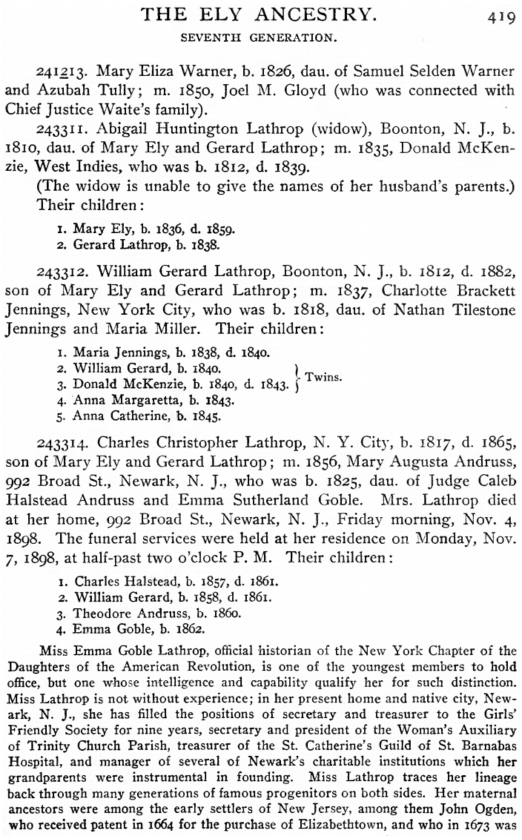
Consider the application of family-history information as captured in historical books. We have access to a collection of 85,000 such books that describe family genealogy, biographies, family stories, photos, and related information. These documents contain a variety of pages as Figures 1 and 2 illustrate. Documents such as these are information-dense, containing many assertions both directly stated and implied. For example, from the page in Figure 1 we read that Mary Ely was born to Abigail Huntington Lathrop in 1838—the author stated this assertion directly. However we also can infer that Mary was a daughter of her mother Abigail because "Mary" and "Abigail" are generally accepted as a female names. This type of information—including both stated and inferred assertions—is useful to someone who is searching for information about members of this family.

Assume that each book has approximately 500 pages, that there are 100 stated and 100 inferred assertions per page, and that each assertion requires 500 bytes of storage. Further assume that each page needs to be stored both as a high-resolution image and as a processed textual representation, taking another 10,000 and 1,000 bytes respectively.[5] We conservatively estimate that we could store a fact base extracted from these 85,000 documents in $85,000 \times 500 \times ((100+100) \times 500 + 10,000 + 1,000) = 4,717,500,000,000$ bytes. The 4.7 terabytes constitutes a modestly large data store, though fairly manageable, and we guess that compression techniques could reduce the storage requirement into the sub-terabyte range.

However, this collection is only the beginning within the family-history application domain. There are many such collections of historical family-history books, and a variety of other related information sources of interest, both static

[4] Moore's Law is not strictly speaking a law, but rather Gordon Moore's observation that the number of transistors on an integrated circuit doubles approximately every two years. The observation has generally held true since 1965, though some observers believe the rate of growth will soon decrease. See `http://en.wikipedia.org/wiki/Moore's_law`.

[5] These assumptions are based on approximate averages we have observed in our actual work on historical documents.



THE ELY ANCESTRY. 419

SEVENTH GENERATION.

241213. Mary Eliza Warner, b. 1826, dau. of Samuel Selden Warner and Azubah Tully; m. 1850, Joel M. Gloyd (who was connected with Chief Justice Waite's family).

243311. Abigail Huntington Lathrop (widow), Boonton, N. J., b. 1810, dau. of Mary Ely and Gerard Lathrop; m. 1835, Donald McKenzie, West Indies, who was b. 1812, d. 1839.

(The widow is unable to give the names of her husband's parents.) Their children:

1. Mary Ely, b. 1836, d. 1859.
2. Gerard Lathrop, b. 1838.

243312. William Gerard Lathrop, Boonton, N. J., b. 1812, d. 1882, son of Mary Ely and Gerard Lathrop; m. 1837, Charlotte Brackett Jennings, New York City, who was b. 1818, dau. of Nathan Tilestone Jennings and Maria Miller. Their children:

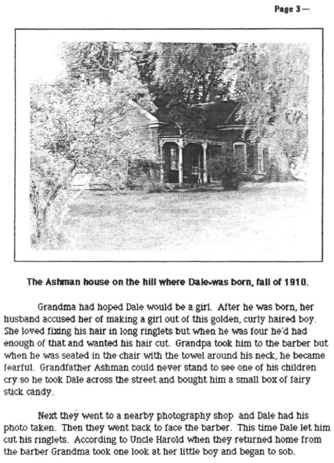
1. Maria Jennings, b. 1838, d. 1840.
2. William Gerard, b. 1840. } Twins.
3. Donald McKenzie, b. 1840, d. 1843. }
4. Anna Margaretta, b. 1843.
5. Anna Catherine, b. 1845.

243314. Charles Christopher Lathrop, N. Y. City, b. 1817, d. 1865, son of Mary Ely and Gerard Lathrop; m. 1856, Mary Augusta Andruss, 992 Broad St., Newark, N. J., who was b. 1825, dau. of Judge Caleb Halstead Andruss and Emma Sutherland Goble. Mrs. Lathrop died at her home, 992 Broad St., Newark, N. J., Friday morning, Nov. 4, 1898. The funeral services were held at her residence on Monday, Nov. 7, 1898, at half-past two o'clock P. M. Their children:

1. Charles Halstead, b. 1857, d. 1861.
2. William Gerard, b. 1858, d. 1861.
3. Theodore Andruss, b. 1860.
4. Emma Goble, b. 1862.

Miss Emma Goble Lathrop, official historian of the New York Chapter of the Daughters of the American Revolution, is one of the youngest members to hold office, but one whose intelligence and capability qualify her for such distinction. Miss Lathrop is not without experience; in her present home and native city, Newark, N. J., she has filled the positions of secretary and treasurer to the Girls' Friendly Society for nine years, secretary and president of the Woman's Auxiliary of Trinity Church Parish, treasurer of the St. Catherine's Guild of St. Barnabas Hospital, and manager of several of Newark's charitable institutions which her grandparents were instrumental in founding. Miss Lathrop traces her lineage back through many generations of famous progenitors on both sides. Her maternal ancestors were among the early settlers of New Jersey, among them John Ogden, who received patent in 1664 for the purchase of Elizabethtown, and who in 1673 was

Fig. 1. *Ely Ancestry* page.



Page 3 —

The Ashman house on the hill where Dale-was born, fall of 1910.

Grandma had hoped Dale would be a girl. After he was born, her husband accused her of making a girl out of this golden, curly haired boy. She loved fixing his hair in long ringlets but when he was four he'd had enough of that and wanted his hair cut. Grandpa took him to the barber but when he was seated in the chair with the towel around his neck, he became fearful. Grandfather Ashman could never stand to see one of his children cry so he took Dale across the street and bought him a small box of fairy stick candy.

Next they went to a nearby photography shop and Dale had his photo taken. Then they went back to face the barber. This time Dale let him cut his ringlets. According to Uncle Harold when they returned home from the barber Grandma took one look at her little boy and began to sob.

Fig. 2. *Dale Ashman* page.

and dynamic. For example, census records, ship manifests, historical newspapers, parish records, and military records are just a few of the types of information that a family-history company like Ancestry.com is interested in gathering, integrating, and making available to its clients. In addition to static sources, there are also dynamic sources such as family blogs, shared photo albums, and the Facebook social graph that could usefully augment the historical document base. Taken together, these sources easily exceed many exabytes of data. So the family-history domain certainly exhibits the *volume*, *variety*, and *velocity* challenges characteristic of Big Data. This domain also expresses the *veracity* dimension: it is common for multiple sources to make conflicting assertions about family-history details such as dates, places, names, person identity, and familial relationships.

Returning now to the relatively modest collection of 85,000 historical books, it is true that a search engine such as Lucene[6] could readily be used to construct a full-text keyword index of this document base. However, keyword search, while

[6] See http://lucene.apache.org.

Ontology Snippets:
ChildRecord
external representation: ^(\d{1,3})\.\s+([A-Z]\w+\s[A-Z]\w+)
(,\sb\.\s([1][6-9]\d\d))?(,\sd\.\s([1][6-9]\d\d))?\.
predicate mappings: *Child*(x); *Child-ChildNr*(x,1); *Person-Name*(x,2);
Person-BirthDate(x,4); *Person-DeathDate*(x,6)

Fig. 3. Ontology Snippet Example. (The **predicate mappings** associate the text recognized by the regular-expression capture groups 1, 2, 4, and 6 with new child x in their respective relationships in Figure 4.)

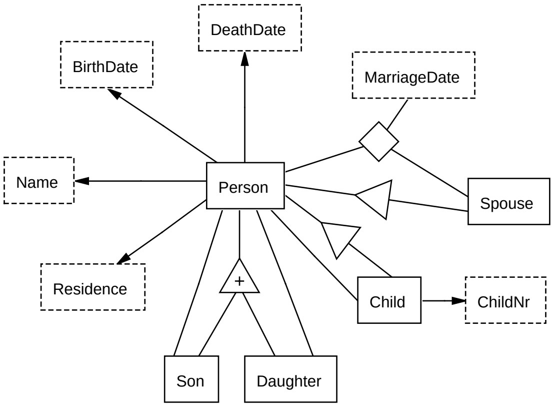


Fig. 4. Ontological Conceptualization for Assertion Extraction.

a good start, is not nearly enough to accomplish the types of semantic searches we need. Is it possible to apply semantic markup to the concepts contained within those pages, semantically index the information for search and query, and clean and organize it as a valuable storage repository? Manually, with crowd-sourcing, this may be possible, but neither the expense nor the timeliness would be tolerable. We see several ways conceptual modeling can "come to the rescue" to enable this application—and, by implication, to enable similar applications:

– *Conceptual-Model-Based Information Extraction.* To address the Big Data issues of *volume*, *velocity*, and *variety*, we take a conceptual-modeling approach to information extraction. We create a conceptual model that conforms to an author's point of view and linguistically ground the conceptual model, turning it into an extraction ontology [ECJ$^+$99,ELLT11]. We linguistically ground a conceptual model by associating regular-expression pattern recognizers with the object and relationship sets of the conceptual model or with coherent collections of object and relationship sets, which we call ontology snippets. For example, we can declare the ontology snippet in Figure 3 to extraction information into the conceptual model in Figure 4, which represents the author's view of the child lists in Figure 1. Further, since manual creation of pattern recognizers is likely to be too expensive for the

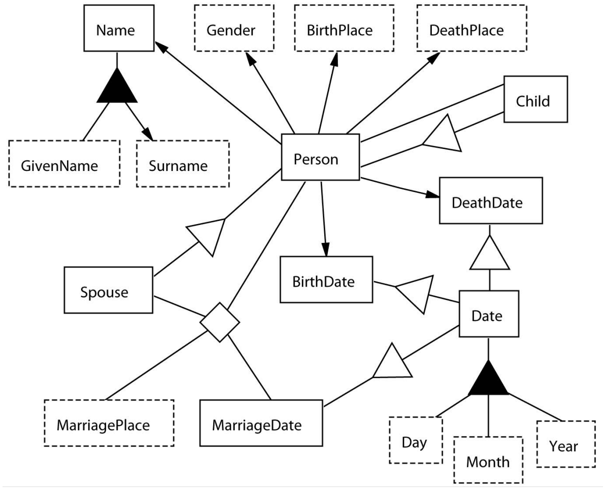


Fig. 5. Target Ontology of Desired Biographical Assertions.

volume, *velocity*, and *variety* of Big Data applications, we seek for ways to automatically generate recognizers (see [PE13a] for an example).

- *Conceptual-Model-Based Knowledge Organization.* Information that is extracted with respect to an author's view is often not ideally organized for search and query. Moreover, we are often interested not only in the stated assertions that can be extracted but also in what can be inferred from the stated assertions. Figure 5 shows a conceptualization of the way we may wish to organize the information in Figure 4 or the information extracted from any other historical document containing family-history information. Because conceptual models are or can be formally based on predicate calculus, we can use inference rules that map from one conceptual model to another to organize our knowledge base. For example, we can reorganize the *Son* and *Daughter* information in Figure 4 as *Child* information in Figure 5 and the *Name* as a multi-token string into an aggregate of *GivenName*s and a *Surname*. We can also infer *Gender*, which is almost never stated explicitly, either from the *Son* and *Daughter* classification or from *GivenName*s based on a probabilistic model of male and female names in nineteenth century America. (See [PE13b] for an explanation about how we use Jena[7] inference rules to map one conceptualization to another.) Besides conceptual organization, we would also like to resolve object identity. Of the four mentions of the name "Mary Ely" in Figure 1, three denote the same person, but the "Mary Ely" who is the daughter of Abigail is certainly different since she is the granddaughter of the other Mary Ely. We take a conceptual-modeling approach to resolving object identity. We extract and organize facts and then check, for example, whether two people with similar names have the same

[7] http://jena.apache.org/

parents or were born in the same location on the same date. (See [PE13b] for an explanation about how we use the Duke[8] entity resolution tool to resolve object identity.)

- *Conceptual-Model-Based Semantic Indexing and Query Processing.* To support the unlocking of the "veritable gold mine of information" in Big Data applications, we provide a conceptual-model-based, semantic-search mechanism that includes semantic indexing, free-form and advanced form-based query processing, and cross-language query processing:
  - Semantic Indexing. To answer queries quickly, we must semantically crawl and index resources in advance. To create semantic indexes, we apply conceptual-model-based extraction ontologies to resources; we also pre-execute inference rules so that we index not only stated assertions but also inferred assertions [EZ10].
  - Free-form Query Processing. We process free-form queries in a hybrid fashion. We first apply extraction ontologies to a query to discover any semantics and leave the rest (minus stopwords) for keyword processing [ELL$^+$12]. For example, for the query "birth date of Abigail, the widow" the extraction ontology in Figure 5, with good recognizers, would discover that "birth date" references the *BirthDate* object set, that "Abigail" is a name in the *GivenName* object set, and that "widow" is a keyword. Hence, the query processing system would generate a query that joins over the relationship sets connecting the identified object sets in Figure 5, selects with the constraint *GivenName = 'Abigail'*, and projects on the mentioned object sets—*Year* of *BirthDate* and *GivenName* for this query. The semantic index links to the pages on which (1) the name "Abigail" and a birth year are mentioned, and (2) the keyword "widow" is present. Since the page in Figure 1 has both, the page-rank algorithm would place it high on its list.
  - Advanced Form-based Query Processing. Because we process queries with extraction ontologies based on conceptual models, once an extraction ontology is identified as being applicable for a query, the system may use it to generate a form for advanced query processing. The query processing system treats all constraints in a free-form query conjunctively, but the generated form allows for the specification of negations and disjunctions as well [ELL$^+$12].
  - Cross-Language Query Processing. Since extraction ontologies are language independent, we can both semantically index and process queries in any language. (In our research we have implemented test cases for English, French, Japanese, and Korean.) We process cross-language queries by requiring that the extraction ontologies for each language have structurally identical conceptual-model instances. Thus, we are able to interpret a query with the extraction ontology in language L_1 and translate the query at the conceptual level to the extraction ontology in language L_2. We can then execute the query over the semantic and keyword in-

[8] http://code.google.com/p/duke/

dexes to obtain a result in language L_2, which can then be translated back into language L_1 for display to the user [ELLT11,ELL$^+$12].

- *Conceptual-Model-Based Constraint Checking.* To address the Big Data issue of *veracity* in our family-history application, we envision applying the constraints declared in a conceptual model to check constraint violations. For example, a person should have only one mother. Because the data is obtained through information extraction and through other means such as crowd sourcing and wiki-like updates by the general public, we allow conflicting information to enter into the system, resulting in a myriad of constraint violations: "I'm my own grandpa", as the saying goes, occurs in the actual (fairly massive) amount of data collected so far [Can13]. Big Data quality [Bat12] will become a huge issue for our family-history application.
- *Conceptual-Model-Based Ontology Construction.* Ontology construction is one of the bottlenecks preventing the semantic web from achieving its envisioned potential as a Big Data application. Conceptual modeling can play a role in automating ontology construction. We have experimented with an approach to automated ontology construction, that takes a collection of tables all on some topic (e.g., *Statistics Canada*, http://www.statcan.gc.ca/start-debut-eng.html), interprets each table and reverse-engineers it into a conceptual-model instance, and integrates the conceptual-model instances into an ontology that covers the concepts and relationships discovered in the collection of tables [TEL$^+$05].

The era of web-scale applications and Big Data is here to stay. As conceptual-modeling researchers we should look for ways to integrate our theory into the practice of Big Data. We see excellent opportunities in all four dimensions of Big Data (*volume*, *velocity*, *variety*, *veracity*) and expect that the community can find more beyond those mentioned here in connection with our efforts to superimpose a web of knowledge over historical documents.

References

[Bat12] C. Batini. Data quality vs big data quality: Similarities and differences. In *Proceedings of the 1st International Workshop on Modeling for Data-Intensive Computing*, Florence, Italy, October 2012.

[Can13] A.B. Cannaday. Solving cycling pedigrees or "loops" by analyzing birth ranges and parent-child relationships. In *Proceedings of the 13th Annual Family History Technology Workshop*, Salt Lake City, Utah, USA, March 2013.

[ECJ$^+$99] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering*, 31(3):227–251, 1999.

[ELL$^+$11] D.W. Embley, S.W. Liddle, D.W. Lonsdale, S. Machado, T. Packer, J. Park, and N. Tate. Enabling search for facts and implied facts in historical documents. In *Proceedings of the International Workshop on Historical Document Imaging and Processing (HIP 2011)*, pages 59–66, Beijing, China, September 2011.

[ELL+12] D.W. Embley, S.W. Liddle, D.W. Lonsdale, J.S. Park, B.-J. Shin, and A. Zitzelberger. Cross-language hybrid keyword and semantic search. In *Proceedings of the 31st International Conference on Conceptual Modeling (ER 2012)*, pages 190–203, Florence, Italy, October 2012.

[ELLT11] D.W. Embley, S.W. Liddle, D.W. Lonsdale, and Y. Tijerino. Multilingual ontologies for cross-language information extraction and semantic search. In *Proceedings of the 30th International Conference on Conceptual Modeling (ER 2011)*, pages 147–160, Brussels, Belgium, October/November 2011.

[EZ10] D.W. Embley and A. Zitzelberger. Theoretical foundations for enabling a web of knowledge. In *Proceedings of the Sixth International Symposium on Foundations of Information and Knowledge Systems (FoIKS'10)*, pages 211–229, Sophia, Bulgaria, February 2010.

[PE13a] T.L. Packer and D.W. Embley. Cost effective ontology population with data from lists in ocred historical documents. In *Proceedings of the 2nd International Workshop on Historical Document Imaginag and Processing (HIP 2013)*, Washington D.C., USA, August 2013. to appear.

[PE13b] J.S. Park and D.W. Embley. Extracting and organizing facts of interest from ocred historical documents. In *Proceedings of the 13th Annual Family History Technology Workshop*, Salt Lake City, Utah, USA, March 2013.

[TEL+05] Y.A. Tijerino, D.W. Embley, D.W. Lonsdale, Y. Ding, and G. Nagy. Toward ontology generation from tables. *World Wide Web: Internet and Web Information Systems*, 8(3):261–285, 2005.

[ZEd+11] P.C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. *Understanding Big Data*. McGraw-Hill, Inc., New York, NY, USA, 2011.