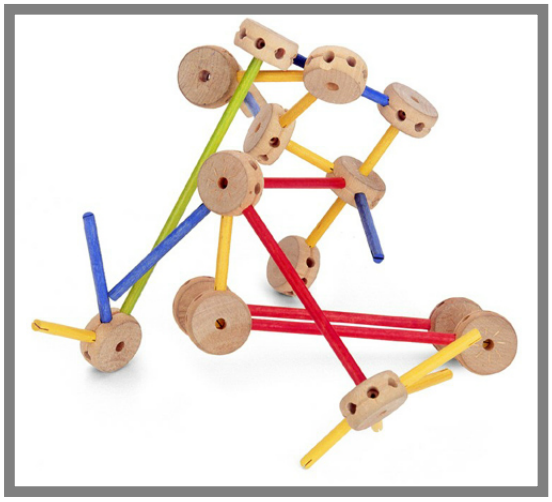
What is Big Structure?

by Mike Bergman - Tuesday, August 12, 2014

http://www.mkbergman.com/1778/what-is-big-structure/



Defining the Guideposts for Big Data

In our recent two-part series we described a decade of experience working in the semantic Web (Part I) and our view that *Big Structure*, which resides at the nexus of the semantic Web, knowledge bases and artificial intelligence, was a key component of making sense of Big Data going forward (Part II). We are at a time when multiple advances are conjoining to create new opportunities and excitement.

Data without context and relationships is meaningless. The idea of Big Data is powerful, but it is often presented as either a "good thing" in and of itself, or a mantra for something that is rather undefined. There is no doubt that with the Internet and the Web we are now able to generate and access data at unprecedented scale. There is also no question that tracking mechanisms and cheap storage -- and simpler, large-scale databases and Web services -- mean that we can also capture data and structure of natures previously unseen. Everyone knows the remarkable growth in exabytes and more.

The prospect of data everywhere -- some useful with important context and some not -- has clearly captured the current discussion. Heck, if we claim Big Data, we even make more in wage or consulting charge-out fees. Who can argue with that?

Well, actually, anyone interested in meaningful data or cross-dataset interoperability can argue with that. Big Data is great, except it means little if we can not combine that data across multiple sources for potentially multiple purposes. (Remember, one of the "V's" of Big Data is variability.) Once the question of what data means gets brought to the fore, it is now time for context and relationships. Structure in an information context means that which situates or describes data in an interpretable way. Big Data needs a Big Structure complement to make sense of it all.

What is a Big Structure?

Big Structure is data relationships and context that can be combined into a coherent framework to enable dataset interoperability and understanding. By necessity, Big Structure implies that the meaning of data can be understood and its values can be brought to common bases such that analysis, testing and validation can be applied across values. Big Structure is not a monolithic thing, but the combination of multiple things that give data meaning and context. As such, Big Structure is often a re-purposing of existing structural assets, plus other special sauce, organized for the aim of data interoperability.

Big Structure is data relationships and context that can be combined into a coherent framework to enable dataset interoperability and understanding.

The components of Big Structure can be identified and characterized. These components can be assessed for usefulness and authoritativeness, and then incorporated into broader structures that ultimately bring the topics of what the data is about and the values of that data into alignment. Thus, Big Structure is also a mindset and approach to selecting and combining structures such that broad dataset interoperability can be achieved.

Big Structure is actually a continuum or family of concept and data relationships, any one of which is also a contributor to helping to map and interoperate data. Ultimately, the components of Big Structure get combined into reference graph structures that place the concepts and actual data values of the Big Data into context. There are certain ways to use and organize existing structures to achieve these Big Structure objectives; some of these ways are described in this article.

Once the components of Big Structure are combined into these reference graphs we then can also use network or graph analysis to understand the relationships amongst the constituent data items. This recursive nature of graph reference structures to organize the constituent data and then to use those graphs to analyze the data is one of the hallmark characteristics of Big Structure.

Big Structure thus involves the need to identify and then organize constituent forms of structure into coherent reference frameworks. Concepts in contributing datasets are then mapped to these structures, and the attributes and values of the underlying data are also transformed into canonical representations. It is these mappings and transformations that provide the interoperability of Big Structure. Big Structure therefore continues to evolve by adding more and more reference structures, all coherently organized.

Contributors to Big Structure

Big Structure is a family of canonical reference structures that help guide mapping and interoperability. The table below lists some of the possible contributors to Big Structure [1], roughly in descending order as to the degree of structure and its contribution to interoperability. The table provides both definitions and use descriptions for each component, plus optionally some notes regarding coverage and use:

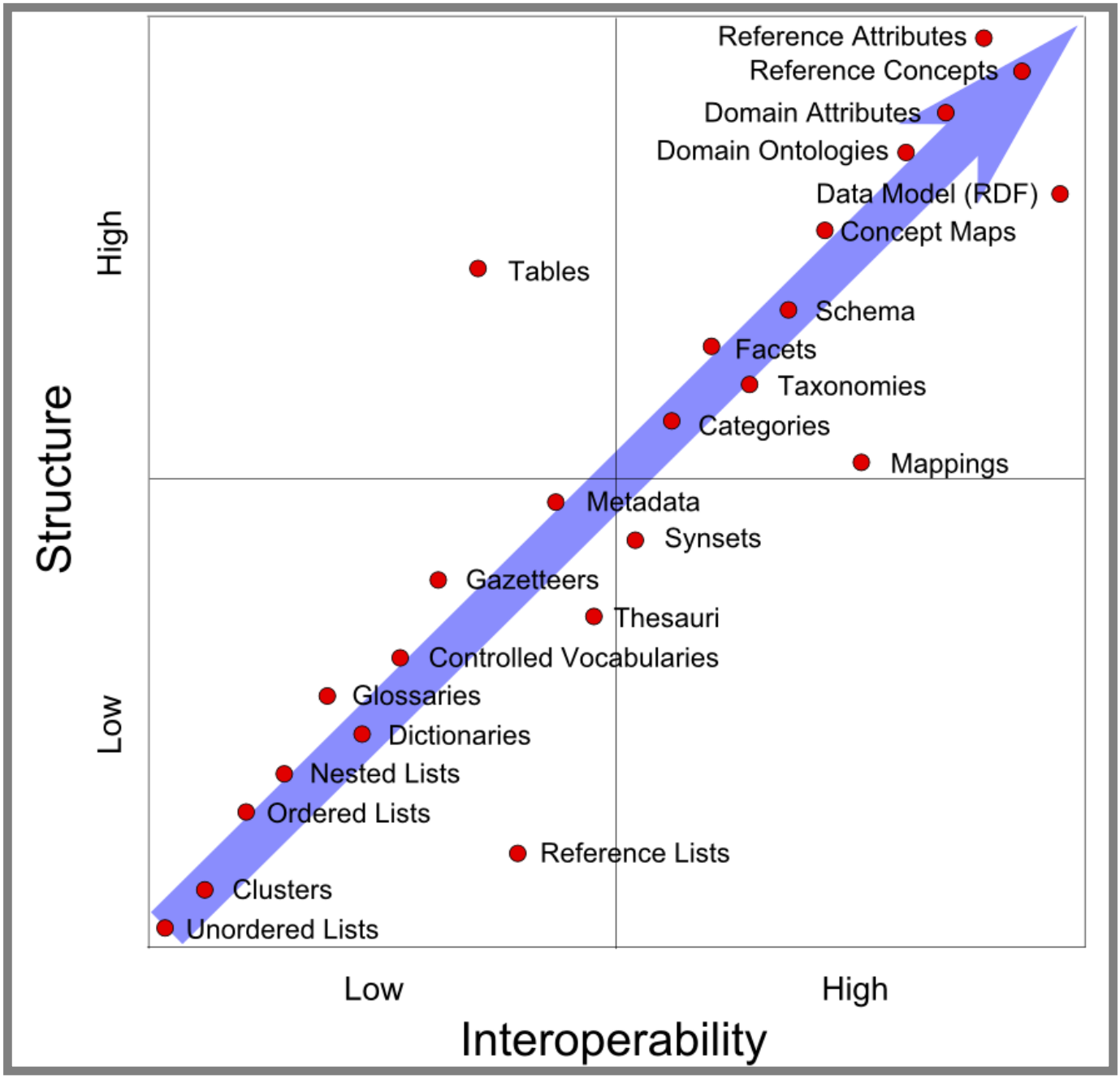
Structure Type	Definition	Use

Reference ontologies	Major grounding structures for orienting and interoperating concepts or data	The reference concepts for orienting all data and domain information
Reference attributes	Major grounding structures for interoperating data and data characterizations	The reference relationships amongst data descriptions and characteristics, which also provides the means for transformations between heterogeneous representations
Data model (RDF)	A self-consistent means for describing the structure of data and their relationships	The "canonical" data model at the heart of the system; provides a single interoperability point; RDF is the canonical model used by Structured Dynamics for its Big Structures
Domain attributes	The data descriptions and characteristics for the constituent datasets in the applicable domain(s)	The reference attributes specific to the domain(s) at hand (which are generally more specific than general reference attributes)
Domain ontologies	The formal conceptualization of a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts	The reference concepts and their relationships specific to the domain(s) at hand; generally are mapped to the reference ontologies
Concept maps	A diagram that depicts suggested relationships between concepts	Structurally similar to a domain ontology; a few related terms shown in Note
Schema	The structure of a database that defines the objects and relationships in that database	Organizing framework for relational databases (and their tables)
Mappings	The process of creating data element correspondences between two distinct data models or schema	Mapping predicates are used to relate concepts or attributes from two different datasets or knowledge bases to one another. Mappings are often a precursor to various transformations to bring data into a common representation
Taxonomies	A particular classification of related concepts,	Hierarchical relationships

	often of a hierarchical nature	are expressed in narrower or broader terms (or `subClassOf`); may also be `see also` relationships
Facets	Clearly defined, mutually exclusive, and collectively exhaustive aspects, properties or characteristics of a class or specific subject	Facets can provide alternative ways for classifying objects beyond a single taxonomy
Categories	Grouping objects based on similar properties	A category may be viewed as equivalent to a concept
Tables	A collection of related data held in a structured format, generally a two-dimensional layout of rows (records) and columns (fields)	Simplest and most common data presentation format
Synsets	A group of data elements or terms that are considered semantically equivalent for the purposes of information retrieval	Also known as a "semset" in the parlance of UMBEL
Metadata	Data providing information about one or more aspects of the source data, thus "data about data"	It is the description of what data is about rather than the values and attributes of the actual data
Thesauri	A form of controlled vocabulary that seeks to dictate semantic manifestations of metadata in the indexing of content objects	A thesaurus is composed a list of words (or terms), a vocabulary for relating these words (or terms) to one another, often hierarchically, and a set of rules on how to use these aspects
Gazetteers	A listing of similar entity types with associated structural data (such as countries and population or standard codes)	Often used in relation to people or place entity types, though any class of entities may have a gazetteer
Controlled vocabularies	The use of predefined, authorized terms as preselected by the sponsor to enforce consistency in terminology	Applied to specific domains or sub-domains, with single controlled vocabularies per official language used
Reference lists	Authoritative listings of similar objects, each uniquely identified by name or code	May be as simple as a comprehensive list of countries with associated ISO codes
Dictionaries	A repository of information about data such as meaning, relationships to other data, origin, usage, or format	In our context, can range from the meaning associated with standard word dictionaries to the

		more formal data dictionary
Glossaries	An alphabetical list of terms in a particular domain with the definitions for those terms	Definition is the only structured information provided
Nested lists	Related concepts or entities organized by some form of hierarchical relationship (narrower, broader, subClassOf, etc.)	Akin to a simple taxonomy
Ordered lists	A finite, ordered collection of values for a given type	May also be additional information linked to the listing
Clusters	A set of objects grouped according to some basis of similarity (type, attributes, or characteristics)	Basis for how the objects got clustered is not always obvious
Unordered lists	A container of similar items or entities, with no implied order or sequence	Also known as a "bag" or "collection"
Values	The actual data; a normal form or a type member	Basic QUDT ontologies could contribute here

An alternate way to look at these contributor structures is to characterize them with respect to degree of structure and degree of contributing to interoperability:


Reference Attributes
Reference Concepts
Domain Attributes
Domain Ontologies
Data Model (RDF)
Concept Maps
Tables
Schema
Facets
Taxonomies
Categories
Mappings
Metadata
Synsets
Gazetteers
Thesauri
Controlled Vocabularies
Glossaries
Dictionaries
Nested Lists
Ordered Lists
Reference Lists
Clusters
Unordered Lists
High
Low
Structure
Low
High
Interoperability

Structure v Interoperability

In general, as might be expected, the greater the degree of structure, the greater its potential contribution to interoperability. The components in the upper right quadrant represent the most structured and interoperable ones. These also conform most to the use of W3C standards for the RDF data model and the OWL ontology languages. Expressions of structure are codified and standardized. Use of best practices also ensures completeness and suitability as reference groundings for interoperability.

The lower left portions of the quadrant represent the least structure and interoperability. However, as standard reference means for characterizing and describing data, even structures in this quadrant can contribute to meeting Big Structure requirements. Tagging of documents (unstructured data) occurs in this less-sophisticated lower left quadrant, but it gives equal footing to 80% of the content that generally resides in text form. (The interoperability system is further enhanced when the basis of the tags is derived from the "semsets" of the reference and domain ontologies, another example of a best practice.)

All of the listed components can thus contribute to Big Structure. However, the completeness of that structure and its usefulness for interoperability increases as one progresses along the blue arrow of the Big Structure continuum. Data interoperability arises from the continued efforts to drive Big Structure to the upper right of this quadrant. As noted, Big Structure is a mindset and process rather than some finite state. As more concepts and attributes get grounded in standard references, the degree of Big Structure (and, thus, data interoperability) continues to increase.

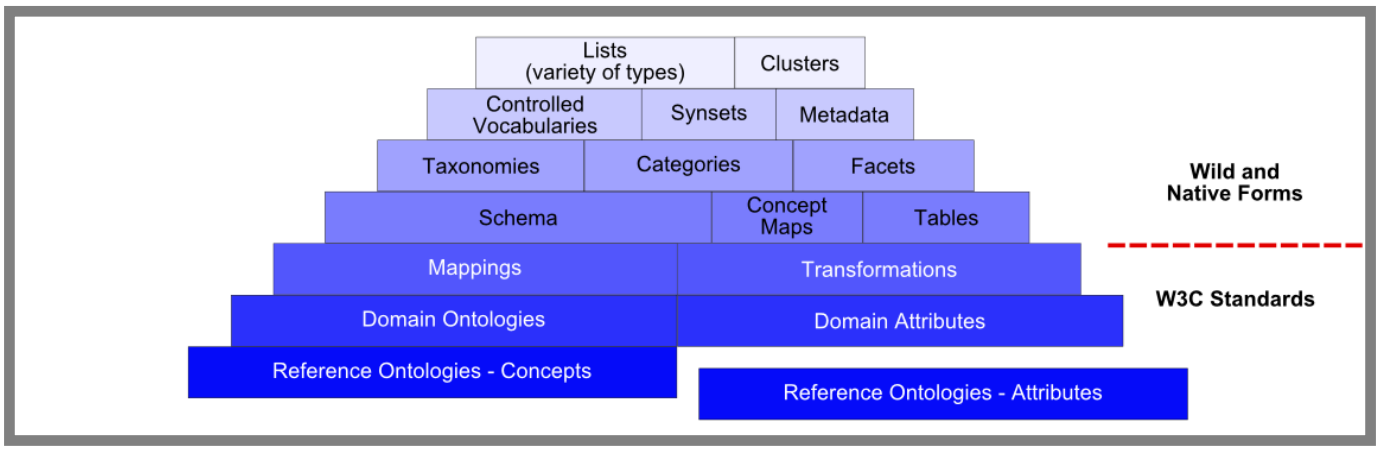
The Foundation of Reference Groundings

In both semantics and artificial intelligence -- and certainly in the realm of data interoperability -- there is always the problem of symbol grounding. In the conceptual realm, symbol grounding means that when we use a term or phrase we are referring to the same thing; that is, the referent is the same. In the data value realm, symbol grounding means that when we refer to an object or a number -- say, the number 4.1 -- we are also referring to the same metric. 4.1 inches is not the same as 4.1 centimeters or 4.1 on the Richter scale, and object names for set member types also have the same challenges of ambiguous semantics as do all other things referred to by language.

The variability V in Big Data or the 40-some dimensions of potential semantic heterogeneity [13] are explicit recognitions of the symbol grounding challenge. Assuming we can determine context (itself an important consideration not further discussed here), fixity of reference is essential to these groundings. Context and groundings are the ways by which we remove ambiguity in what we measure and record.

Like dictionaries for human languages, or stars and constellations for navigators, or agreed standards in measurement, or the Greenwich meridian for timekeepers, fixed references are needed to orient and "ground" each new dataset over which we attempt to integrate. Without such fixities of reference, everything floats in reference to other things, the cursed "rubber ruler" phenomenon.

Thus, we can express our Big Structure components from a foundational perspective as well. In Structured Dynamics' view of the world, the foundation for data interoperability is grounded in reference structures or ontologies that provide the fixity of reference for concepts and data and their attributes. Upon these foundations are then constructed the domain views of concepts and attributes, which become the target for mapping other references and Big Structures:



Foundations to Big Structure

The mappings, transformations and domain and reference ontologies are themselves written in the OWL languages of the W3C and the standards of the RDF data model. At this most expressive end of Big Structure, the representations are in the form of graphs. Network and graph analytics will expand still further business intelligence prospects. The use of these standards with common and testable logic is another means to ensure coherency and interoperability of the Big Structure that results.

Note a key aspect of the grounding foundation is missing: one or more reference ontologies for attributes. Though many examples exist on the concept side, little has been done to explicitly address the questions of data value interoperability. This major gap is a current emphasis of Structured Dynamics, with much that will be said over the coming weeks. Also expect an open source reference ontology for attributes in the near future.

The thing is that we are learning how to make the various parts of this interoperability stack work. We are leveraging existing structural assets of all kinds to establish the semantics and infrastructure for domain interoperability. We know how to match and map these existing structural assets to the reference frameworks that are the foundation to interoperability.

A Vision of Interoperability

The real world is one of heterogeneous datasets, multiple schema and differing viewpoints. Even within single enterprises -- and those which formerly expressed little need or interest to interoperate with the broader world -- data integration and interoperability has been a real challenge. Big Data itself is not solving these problems. Quite the opposite. Big Data trends are turning data interoperability molehills

into mountain-high competitive threats.

Like any well-built structure, data interoperability requires a solid foundation. That foundation must reside in exemplar reference ontologies upon which to ground the semantics and exchange standards for data. Using the canonical RDF data model makes this task practical. Existing information structures of various types across the enterprise and the Web all can and should play a role in establishing reference structures. The accretion of reference structures will lead to still further interoperability and the ability to incorporate more datasets. Currently expensive practices in, say, master data management (MDM) can begin to transition to a new paradigm. It is easy to envision working from a library of existing reference standards for use across enterprises. This kind of incremental expansion of interoperability leads to still more interoperable data in a virtuous cycle of innovation and lower budgets.

As our computing continues to get more virtual and cloud-like, physical and hardware and software architectures must give way to information architectures (in the true sense of interoperability). We have no choice but to treat the architecting of information as a first-order challenge. The totally cool thing about the data integration challenge is that the architecture can be readily varied and tested to achieve a working foundation. Much empirical information exists about how to do it and what to do next. The chief challenge has been to recognize that data interoperability -- and its dependence on Big Structure -- is a first-order concern (and opportunity). The intersection of Big Structure with Big Data, and with graph and AI algorithms, should create new approaches to chew across the data integration environment. I expect progress to be rapid.

---

[1] There are at least 40 terms or concepts across these various disciplines, most related to Web and general knowledge content, that have organizational or classificatory aspects that — loosely defined — could be called an "ontology" framework or approach. See M.K. Bergman, 2007. An Intrepid Guide to Ontologies, *AI3:::Adaptive Information* blog, May 16, 2007.

[2] UMBEL and other upper level ontologies are examples here. In the case of UMBEL, that Big Structure is used as a scaffolding of reference concepts used to link external (unrelated) structures to help inter-operating data between two unrelated systems. Such a Big Structure can also be used for other tasks such as helping machine learning techniques to categorize and disambiguate pieces of data by leveraring such a structure of types.

[3] Unfortunately, no reference structures for attributes yet exist. For a discussion of this status, see the thread on the W3C semantic web mailing list beginning at http://lists.w3.org/Archives/Public/semantic-web/2014Jul/0129.html.

[4] Data models encompass a rather broad span. The RDF discussion represents a more formal end of the data model spectrum, wherein there is complete logic, syntax and serialization discussions, more involved than most data models.

[5] Domain ontologies represent the most closely-aligned view of the domain and its relationships of all of the component structures listed.

[6] Concept maps are very closely related to ontologies, and may include topic maps, mind maps and other graph-like structures of concepts.

[7] Schema may apply to many realms, but in the IT and software context schema mostly refers to database schema related to relational databases. These are often expresssed in UML diagrams or XML schema.

[8] Mappings and transformatons are a huge area of diverse structure and different serializations and specifications. Fortunately, the task of mapping external structure to RDF removes the many-to-many issues with most transformation approaches.

[9] Taxonomies mask an entire sub-categories of directories, folksonomies, subject trees, and more. The key aspect is that relevant concepts are expressed in a graph relationship manner to other concepts, often in a hierarchical fashion.

[10] Categories also includes the general classification process.

[11] I would consider a canonical references listing of country names and codes to be a part of Big Structure, since they act as a controlled vocabulary.

[12] This is a key area for including unstructured documents, since tags are a primary means of adding metadata to a document. When the pool of tags is based on the governing reference and domain ontologies, then interoperability is further promoted.

[13] M.K. Bergman, 2006. Sources and Classification of Semantic Heterogeneities, *AI3:::Adaptive Information* blog, June 6, 2006.
