Informational Facts and the Metainformation Inherent in IFacts: the Soul of Data Sciences

Journal of Library Metadata, 2013, 13(1), pages 59-74

Robert M. Losee
SILS, UNC, Chapel Hill, NC 27599-3360
losee@unc.edu

Abstract

Books, archives, and media contain information and meaning, and much of what librarians and information scientists focus on are these information vessels. Metadata is normally assigned to such vessels to assist in organizing and retrieving the information vessels. Facts are often derived from these information vessels through text mining, but facts may be directly produced from observations. A particular kind of fact, an informational fact (IFact), describes the information from the observation of an information producing process. Such IFacts will become a major part of future data libraries. Metainformation, a kind of metadata, is associated with these IFacts, describing the empirical characteristics of the observation that produced the IFact. This metainformation is inherent in informational facts, and is critical in the use of scientific datasets and "big data." The differences between regular facts and informational facts are examined, as well as between metadata and metainformation.

Introduction

While much of library and information science focuses on books, articles, images, music, and archival materials, an additional focus on the underlying components of these materials is useful. This includes the study of data extracted from a variety of sources. Media may be decomposed into different forms, but one form, informational facts, are in many situations superior to other kinds of facts. Using this form of informational fact instead of the larger media vessels, such as books, can result in far more accurate reasoning by computers and people using formally describable systems. Informative facts may be obtained from observations in science or business. Informational facts also support true information retrieval, instead of the bibliographic retrieval that is most often presented as information retrieval.

Consider scientific observations of two objects that appear "red." Assume that one was white, with a red light on it when it was observed, and the other object was actually red. What incorrect scientific conclusions might be made by assuming that the two objects are equivalent? To avoid this sort of problem, information about the observations, including lighting, and so forth, might be retained with the facts about the objects, so that the lighting used with the observation might be considered in determining whether the two objects are the same color. We refer below to this *information about informational facts* as *metainformation*, a kind of metadata, and consider below the problem of science with and without such metadata and metainformation.

Metadata “addresses data attributes that describe, provide context, indicate the quality, or document other object (or data) characteristics” (Greenberg, 2005). Metadata and index features have taken a number of forms in topical systems, often based upon naturalness considerations (Foskett, 1996). While some metadata is topical, metadata may also serve to capture other aspects of the item associated with the metadata, such as title and authorship. Metadata may refer to more traditional library and archival materials, or with more recent emphases, on datasets (Greenberg, 2009).

Metadata often provides a limited description of the entity to which it refers. Many library catalogs show a few subject terms for each item in the library. While there are tradeoffs one can make between phrase length and the number of phrases (Losee, 2004), the overall exhaustivity of the ideas in an entity that one wishes to cover with metadata or metainformation is a value that can be manipulated to provide the desired degree of coverage. When the exhaustivity is small, errors in processing of vaguely defined data will occur.

Metainformation is information about informational facts, generated at the time of the production of the original information, whereas metadata can often be generated at a later time. Metainformation is usually a critical component of informational facts, and the presence and use of metainformation is important for the best use of informational facts. Determining whether to combine informational facts, for example, can best be accomplished if the source of the facts and the nature of the information-producing observations is available in the metainformation. Accurate metainformation cannot usually be supplied by people other than the producer of an informational fact.

Facts are representations of other phenomena and states of the universe. The statement ``The wall is painted white” is a fact about a particular wall. The more abstract fact, ``2 + 3 = 5” appears true, with philosophers disagreeing about which world these mathematical symbols, facts, and truths refer to. Facts have a much finer granularity than do most information vessels, and informational facts are yet more precise than regular facts that do not meet the constraints of informational facts.

Information scientists and data scientists study fact-based systems, emphasizing their use and manipulation. Individual informational facts serve as the most basic representation of information and the most basic, precise form of a fact. The study of “big data” requires an appreciation of the differences between regular facts and informational facts. We also argue below that there is a distinction between what we call metadata and metainformation, associated respectively with regular facts and informational facts. It is argued here that the study of informational facts, *en masse* as well as individually, is needed if we are to fully understand how metadata and metainformation are applied to the data sciences. For example, multiple informational facts with the same metainformation, i.e., the same contexts, can be combined into larger informational facts.

The information obtained by observing individuals and nature may be used in recording sounds, images, and data. This information enters into processes whose output can remain fixed over time for recording purposes. With the development of computers and associated mass storage, large databases have been implemented, with access to some of these datasets being through the Internet. Tables in documents or databases are easily decomposed into facts, usually being decomposed easier than when extracting facts from books, artwork, and so forth. Facts are included in information vessels, but humans usually need to exert effort to extract facts from the media they observe, and the extracting process often produces errors. For example, the painting Mona Lisa has a facial expression that is ambiguous, just as the sentence “I saw her duck” can mean different things. Facts are useful and can serve as the major component of databases, as well as serving as resources for library patrons, who often use a library’s almanacs and data sources.

Information vessels, such as books, videos, and media in general, may be decomposed into facts using any of a range of techniques. Indexing techniques may locate objects for use in other aspects of text mining. Text mining also may be used to learn enough about individual documents or facts to determine relationships between documents or facts. Text mining may also be able to produce summaries of documents, as well as assigning labels of topicality and extracting facts from text and media.

Producing and Manipulating Facts

Information vessels, such as books, videos, and artwork, provide a great deal of information in a single vessel, with metadata being about various aspects of the vessel. Metadata and metainformation are about objects or entities. Facts are simple statements that are held to be true, with metadata describing aspects of the facts. They describe states of the world and can be used to store information about a state of the world, as well as in reasoning about the world and its representations. Logic may be understood as the discipline associated with manipulating factual statements, often in the form of propositions or predicates (Kneale & Kneale, 1984). However, deduction and inference may be misused. Many people have assumed that all swans are white because all the swans *they* saw were white; however there are black swans in some geographic areas of the world. Poor inference may lead some children to believe that all birds can fly over a distance, without considering that most ostriches and penguins do not fly much. Statements and facts themselves, and the relationships between objects, may be described and manipulated probabilistically.

Wittgenstein suggests that basic facts describe links between real things. The *Tractatus* suggests that there are elementary sentences, with operators such as *and*, *or*, and *not* being outside the sentences, serving as connectors. The links between elements of sentences are to real objects, with the words acting as representations, such as a picture (Wittgenstein, 1981). These facts based on empirical evidence serve as the basis for informational facts.

Statements are described by both their syntax, the order in which tokens are arranged in a linear order, and semantics, the meaning or intent associated with using the symbols and tokens in the statement. Some statements, such as “the sky is blue,” might be empirically judged to be true or false. Other orderings of terms, such as “the King of France is bald” or “Colorless green ideas sleep furiously,” may be treated as though they have no meaning and thus no truth value. Because there is no King of France, the first statement has no truth value, and the second sentence has no meaning and no truth value, although the sentence appears on one level to be syntactically correct. Other statements, such as “Qyxi epxo zmbqlb” clearly have no meaning or correctness at any level, and thus there is no discernible meaning and thus no truth value. Clearly, the latter statement has less meaning than the earlier statement about the King of France.

Information may be defined as the characteristics of what is produced by a process at the output of the process (Losee, 2012). A singer may sing a song written on paper or that is stored in their memory, or a painter may capture the beauty of a sunset. The information at the output of these artistic processes acts as a representation which is not an exact copy of the input that the process accepts and manipulates, but can be viewed as being related to the input, thus serving as an (often incomplete and often ambiguous) representation of the input to the artistic process. One may treat the input as *x* and the processing that takes place as a function, *f()*, with the characteristics of the output as a function of the input, *f(x)*.

Statements serve as representations of the processes that produced the representations, as well as serving as representations of the inputs to the process. Each representation can serve as a snapshot of the process and its input, and, as with all snapshots, there may be aspects of the subject that are not captured in an image. An image of a person's face, for example, seldom shows the back of the head. A special case of representation is provided by a communication channel. The input to a communication channel is provided by a transmitting device, which places the messages on the channel.   The output of the channel process occurs at the receiver. Communication exists through the execution of a process, with a high degree of agreement between some of the information in the output and some of the information in the input. In the case of natural language statements, the processes that produce the statements are usually in the human brain, although there are computational algorithms that can be used to produce linguistic statements that are often similar to those produced by a human.

Facts have been used for millennia in describing the world and as a basic concept in philosophy. Facts are usually treated as statements that are true, and thus a fact is expected to be an accurate description of the world and as a building block that represents an aspect of the world. They may be generated from raw observations or they may be extracted from text, for example (Siefkes & Siniakov, 2005).

Facts have been used in the artificial intelligence community for decades, supporting knowledge representation and computer reasoning. Members of the MIT Artificial Intelligence Lab, particularly Terry Winograd, addressed the relationships between objects in an artificial world and the statements needed to describe and manipulate the contents of blocks in an artificial world (Winograd, 1972).

As the study of information retrieval systems and computational linguistics grew, more manipulation of text led to more sophisticated fact extraction techniques in the text processing community. Two basic steps are needed, with sentences first being parsed and the subject, verb, and object being placed in a standard form. In the prose "Bill stared at the sunset. It was beautiful," the sentence "It was beautiful" could have "It was beautiful" replaced with "The sunset was beautiful." This involves the resolution of anaphora, so that pronouns in a sentence, such as "It was beautiful," are linked to the "sunset" that was the subject of a previous sentence (Brassell, 2000). Text extraction has continued to mature both theoretically and practically.

Some facts are placed together into complex entities, sometimes referred to as *frames* (Sowa, 2000). A frame may represent the information available associated with a particular context. A frame about eating in a restaurant might have slots that can be filled, describing the items on the menu, a knife, a fork, water served with the meal, and so forth. An attack frame might have slots for the attacker, the victim, the weapon, and so forth. Note that a knife has different functions in different frames, with a knife in a restaurant frame being for cutting up food, and a knife in an attack frame being used as a weapon. A class of objects may have a ranked list of class attributes, which is very similar to a frame (Pasca, 2007) (Weikum & Theobald, 2010). The relationships may be derived from an ontology (Karkaletsis, Fragkou, Petasis, & Iosif, 2011). Different methods for filling out frames, such as biographical information, have been extensively studied (Garera & Yarowsky, 2009). These generated frames contain many facts that can be used in other forms of analysis.

While frames may be used to capture all the contextual information described in a prose passage, software may also attempt to extract the most important facts in prose, such as the subject of a news story (Kastner & Monz, 2009). Extracting names for key objects may be used in reasoning systems, which may use automatic indexing systems to find the items that compose facts (Willis & Losee, 2013).

Regular Facts and Predicates

Regular facts (RFacts) describe aspects of the world by representing the world using simple statements. For example, stating that the capital of France is Paris is a statement containing a fact expressed in prose. This fact describes a relationship between France and Paris, this relationship being one of "capitalness."

Below we illustrate the discussion about facts by emphasizing a few specific predicates and processes. Studying relationships between items by making rigorous statements about specific kinds of predicates can often lead to a deeper understanding of the relationships. For example, metadata and index terms are usually considered superior if the matching between the terms is "equal to" the item being described rather than if the terms are "not equal to" the item being described.

A fact may be denoted with a predicate, including possibly the name of the predicate (or sometimes the nature of the predicate), and the input and other parameters for the predicate. Notationally, we might refer to the specific use of the addition function and predicate that adds *2+3* as *<5, +, {2, 3}>*. Here the inputs to the process are the set of numbers *2* and *3*, the processing is defined as addition, which we assume is denoted by the plus sign, and the informational output from the process is *5*, the sum of *2* and *3*. We denote the output, the process, then the input, written from left to right, with processing or "flow" moving from right to left. When the input to a process or predicate is a multivalued set of objects, the set is denoted within {braces}. We here have a regular fact that 2+3 = 5. There is an assumption in the common language notion of a fact that all regular facts are true.

As an example of the use of predicates, one class of facts that is popular uses the "IsA" predicate that implies a level of representation. When one states that "x IsA y," this means that *x* is a member of the class of *y*'s. For example, the fact that "the orange on the counter IsA fruit" means that the orange on the counter is a member of the class of fruits. “The orange on the counter IsA snack” means that the orange on the counter is also a member of the class of snacks. Many other predicates and processes are available. These include InstanceOf, which captures whether one operand is an instance of the other, capturing class membership, or the PartOf predicate, which captures components, such as whether the stomach is part of the body. We might thus have the regular fact that the stomach is part of the body.

An important predicate that serves as a common process instrumental within RFacts is Equivalence. An *Equivalence[]* function determines whether two operands are equivalent. What does equality mean in an *Equivalence[]* predicate or function? Does it mean “of the same quantity”? Does it mean that they are the same object? That they have all the same characteristics? The *Equivalence[]* function of the value *5* and the value of *(3+2)* is true in some circumstances and false in others. The text statement “5” is not the same as the statement “3+2”. The arithmetic value of *3+2* and the arithmetic value of *4+1* are equivalent. However, the value of the statements “3+2” and “4+1” are not the same. *Equivalence[y, f, x],* where *y = f(x),* might be applied to the *5=3+2* problem as *Equivalence[5, +, {3,2}].* The semantics of the *Equivalence[]* function may be defined so that the value of two operands are compared syntactically, in which case the statement “2+3” is not equivalent to the statement “5”, or they may be compared arithmetically, so that the value of *(2+3)* is equivalent to the value *5*. Empirical equivalence, *Equivalence*$_E$*[j, k]* is computed as whether all the observable characteristics of *j* and all the observable characteristics of *k* have the same value. The statement *x=3* can be an empirical statement whose truth can be empirically determined, whereas a statement like *y = y* often depends on the semantics of the language and symbols, such as “=” and *y*. Algebraic equivalence may be more empirical, while logical equivalence may be more abstract, and perhaps not empirical at all. We assume that objects are considered equivalent if all observable characteristics are the same. There are actual data and processes, and there are the observed data and observed processes, with the observation of reality being by a person or mechanism.

In summary, a *regular fact*, an RFact, has a predicate and the associated data (or null values) presented to the predicate. For example, the predicate could be *IsA[]* or an *Equivalence*[] in a regular fact.

IFacts

An informational fact (IFact) represents a relationship between empirical observations of the world, or their derivatives. For example, consider a doorbell system that determines whether the button outside the house, near the door, is pressed, and, when the button is pressed, the bell inside the house rings. Here the input to the doorbell ringing process is whether the button outside is pressed or not. The ringing bell represents a combination of information about the doorbell system as well as about whether the button outside is pressed. If there is no ringing, perhaps the button outside is in the “unpressed” state, or the electric power may not be available to the doorbell system, or the system may be “broken.” We will denote such a fact using predicate notation, with the informational fact *About[f(x), f(), x]* representing this case as *About[ringing, bell system, bell pressed]*. The process *f()* in an informational environment produces a representation at its output of the arguments of the process, as well as a representation of the process itself. The About predicate is at the core of any informational fact.

Informational facts are limited to representations of the actual world. This has the potential to exclude “abstract” ideas, such as that zero plus one equals one. Such a mathematical idea or expression may be inferred or deduced from empirical observations but, because it isn’t the result of direct observation, it may not be an IFact. There may be abstract concepts such as zero and one such that the sum of the two would produce one. Instead, one might have learned empirically that individual objects retain the same cardinality or set size, and that adding no object to a container with one object empirically always produces a set of objects of the same size, one. Similarly, the idea of a unicorn may be a combination of empirical observations of a horse or similar actual creature and an actual horn.

The truth of an empirical IFact exists when all the f, x, and $f(x)$ values are observed values, and the observed values all match with the actual values, so that $x_{obs} = x$, $f_{obs} = f$, and $(f(x))_{obs} = f(x)$. When these equalities are true, $About_{obs}[f(x), f, x] = About[f(x), f, x]$. The truth of an empirical IFact may also be computed as $Equivalence_L[True,About[]]$. The truth of a regular fact would be $Equivalence_L[True, RFact[]$. Here logical equivalence $Equivalence_L$ is the logical equivalence between two objects, that they share the same truth value, and $Equivalence_E$ is the empirical equivalence between two objects, where the values of the observable characteristics are the same.

Something can be a non-informational fact, and be a regular fact, in a number of circumstances. For example, if one observed Alice reading the book and saying as a result “that was good,” one can accept the IFact *<”that was good”, “Alice reading and saying as a result”, “the book*>. The statement “that was good” is about Alice reading and speaking and “the book.” One can be fooled, such as if Alice wasn’t actually reading but had an earphone in her ear that was playing an audio recording of the book, or if the entire image of the book was a video seen by the observer that had been generated by a computer. One can incorrectly believe that a true statement is an IFact when it is true but is not based upon empirical observations or their derivatives, making it an RFact.

Other Types of Processes and Predicates

There are a variety of commonly observed representation processes in the information sciences. One representation process makes a perfect copy, the exact reproduction of the input at the output of the process. In arithmetic, an identity function can be used to multiply any input number by *1* to obtain the input number as the output of the process, just as one can add *0* to any input to get the same number at the output. The perfect copy function similarly produces the input at the output.

A representation is like a copy but it contains some features that the observer would find useful for representing the object being represented. One might find a visual representation useful for remembering what someone looks like, that they were tall or short, or what kind of hair the person had. An audio description of the object might remind one of the speech accent of the person represented. A perfect representation accurately and completely captures all observable aspects of the object being represented. This may be treated as a perfect copy.

In most cases, only a partial representation is produced. In this case, the informative output is a subset of the output of a process that produces a perfect representation. Viewing a white building through red tinted glass will make most of the characteristics appear reddish, so that having an image of a reddish door would serve as a partial representation, being a subset of the characteristics of the viewing process. Viewing the house through opaque glass would produce no representation of the house at all.

Reasoning can occur using a number of different kinds of processes. An input or an output to a process may be very elementary or atomistic, such as the frequency and amplitude of a sound at a certain brief time, or a higher level process may produce human thoughts, using many subprocesses, which may, in turn, have their own set of subprocesses, and so forth. The informative output from a process may be on any of a number of levels, ranging from the atomic level to higher level, and from simple to more complex processes.

Facts can exist at a number of levels, with the inputs, outputs, and processes themselves being of any desired granularity. They may represent completely or poorly. Facts associated with higher level objects and processes may be dependent upon the existence and truth of lower level, finer granularity objects and processes. The statement that the "cat is white" can serve as the basis for a single fact, such as <*white, furcolor, cat*>, or it can be decomposed into a set of individual facts about each hair, or facts may be used at a lower level about the genes in each cell in each piece of hair. The original, high level process may be examined as presented, or viewed in a decomposed set of processes along with the initial input. A hierarchy of processes may also be seen as a series of processing units between input and output (Losee, 2012). Informative processes may be deterministic, where the same output is always obtained from a given input. Processes may also be probabilistic, with the output varying given a specific input.

"About" Predicates in Facts compared to other Predicates

Informational facts (IFacts) are always empirically supported, while many other kinds of facts, such as those based on equivalence or IsA relationships, often are not, often being based upon assumptions. These IFacts can be used to study the world scientifically. Scientists often attempt to replicate scientific facts, with repeated processes, inputs, and outputs. An informational fact represents both the data and the information producing process associated with manipulating the data. Based on actual observations instead of on suppositions, an individual IFact should be as accurate or more accurate than other kinds of facts. The standards for what constitutes a representation is much lower than for equivalence. The Mona Lisa represents Lisa Gherardini. One cannot correctly state that the painting Mona Lisa is identical to Lisa Gherardini, just as the words written by the author are not equivalent to the author's ideas. One can say with complete accuracy that the painting Mona Lisa is *about* Lisa Gherardini and the painter DaVinci.

The *About[]* function may be empirically defined to describe operations in the real world. By noting the input to the process, the output to the process, and the process itself, one can precisely capture the state of the real world with an informational fact, with the precision being captured by metainformation about the IFact. If those with accurate facts will better model their world and actions and will perform better, choosing to use those facts that can most accurately reflect reality will result in the most useful facts.

Equivalence[] may take on several different semantic forms, but these do not always accurately reflect reality. Saying that two kittens "are the same" does not reflect that either of the kittens have all the same characteristics, or that they are actually the same thing. Gathering data may produce strong evidence that *x* is *y,* but it does not show equivalence of *x* and *y*. One might conclude that *x* is equivalent to *y* with a high probability, but this is not unequivocal equivalence. Mathematical proofs may argue on theoretical grounds that, for example, *1 = 1,* so that the statement *Equivalence[1,1]* could be known to be true.

If there are advantages to using the *About[]* process or predicate instead of a non-empirical predicate, such as *Equivalence[],* then the informative IFacts may be more precise than regular RFacts or, at worst case, the IFacts will be as strong as RFacts.

An IsA function is based upon a limited amount of empirical support, because of the limits of empiricism. The fact that the orange on the counter is a fruit may be empirically verifiable, but facts such as that all oranges are fruits may be true either by definition or because one has examined all the oranges that exist and thus one can conclude that all are fruits. One could extrapolate from the fact that the orange on the counter has all of a set of characteristics that are used to define fruits so that, in fact, the orange on the counter is a fruit. However, scientists could be wrong about oranges; perhaps there will be an orange that grows as a nut or as a vegetable. Swans were once thought by many to be all white, but non-white swans were later discovered.

Using IsA functions can have further theoretical complications, as suggested by Russell's Paradox. This paradox suggests that if one assumes that *x* is the set of all sets that are not members of themselves, and if *x* is a member of itself, then a contradiction exists. If *x* is not a member of itself, then it would qualify as a member of the set *x* by definition.

About[] more accurately describes the world than *IsA[]* in facts, and *About[]* is more precise and more informative than *Equivalence[]* in IFacts. IFacts using *About[]* predicates will generally be more precise than RFacts.

Inherent Metainformation in IFacts

Metainformation is commonly found in informational facts and is inherent in most, if not all forms of informational facts. Metainformation is information about information. We might view metainformation as "attached" in some sense to the original information from an observation function whose output becomes a component in an IFact. These words could have attached to them the metainformation about who produced them (the author). Metainformation describes the environment in which information develops and occurs, and this information is used when describing the observations that exist in IFacts, as well as combining multiple IFacts.

IFacts use metainformation to represent characteristics of observations from reality. For example, one might state that "I observed on July 15 2007 that the bird feather was blue that I found on the ground at 9:00 AM at Geographic Location 35.91N

79.05W." One can have a basic fact that <*blue, observation, bird feather*> with attached metainformation about the time and the place of observation, as well as who observed the feature. Any observation takes place in a context, and thus any empirical informational fact will have inherent in the fact some metainformation about the observational act. Note that this metainformation may be attached to the IFact or it may not be attached.

We might treat the metainformation associated with an IFact as being either a representation for an observed object or the characteristics associated with the observation that takes place. Several different informational facts with their associated metainformation may be combined and generally manipulated using a number of metainformation operators to combine the IFacts themselves, possibly producing a more panoramic IFact. The combination of IFacts is often dependent on there being similar or identical metainformation for each of the IFacts.

An observation occurs within a particular context. One characteristic of a fact is who conducted the observation of the fact. This might be a name or representation of the observer. The position of the observer serves as an important type of context, with observations having a geometric aspect, as one looks in a certain direction to observe something in the room in which one is situated. Some characteristics of the observation include (1) location of observer (2) time of observation (3) directions in space of observations, (4) transformations that occur within the observing system, and (5) scope or nature of observation. When the positional metainformation for two observations is compared and found to be present and equal, then the observations may be combined by placing these observations together, such as with concatenation.

Metainformation within IFacts about observed objects may be combined to produce new, empirically justified facts, which are "anded" or concatenated to produce compound facts. Rules for combining IFacts based upon similar metainformation will often be context dependent. When IFacts are combined, attention can be paid to the metainformation available about each IFact to determine how and whether the facts should be combined, and if so, how. Reasoning with accurate information about the origin (or nuances) of the IFacts will produce more accurate results as the results are combined with other results as conclusions are used for other reasoning operations.

With the "objects" above, what is observed or who is the observer may be observed and described with facts. These may, in turn, be described in terms of other facts and objects, and so forth. An infinite regression is possible, and an extremely large, and possibly infinite amount of metainformation may be used in any informational fact. This suggests that there may never be pure IFacts, but instead, approximations to the true informational facts when production of the IFacts is truncated. We treat IFacts as having only one level of representation, and for our purposes, they do not regress forever. Some representations may have higher resolution than others, and the higher resolution representations may be used in lieu of the lower resolution representation.

The metainformation in IFacts is different than the metadata used to describe other entities, both *realia* and bibliographic entities. Traditional metadata describes a range of characteristics of the object of its description, from the authorship to topical information (Greenberg, 2009). Metadata may be added by information specialists to support improved retrieval of the facts. For example, a fact that refers to "dog" might be more likely to be retrieved if the topical term "canine" were also added to the fact as metadata. The metainformation associated with IFacts must be obtained, directly or indirectly, from the actual process that produced the IFact. Traditional metadata, on the other hand, may be produced long after the information production process, based on the characteristics of the fact, such as what are currently perceived to be the topics covered in a document. Metainformation contains information that may not always be inferred at a later time, such as the time at which the data was observed.

Relative Amounts of Information Present in RFacts & IFacts

A statement in prose, such as “On an average, men weigh more than women,” represents an idea in the speaker’s mind, given an understanding of some information present about men, women, weights, and averages. A set of regular facts, such as, “Alice’s weight is 50 kilograms. Bob’s weight is 70 kilograms. Charlene’s weight is 48 kilograms. Doug’s weight is 75 kilograms. Ellen’s weight is 60 kilograms. Fred’s weight is 80 kilograms,” along with facts about the sex of various people and knowledge of an algorithm for computing averages, would allow one to utter a prose statement such as “on an average, men weigh more than women.” There is more data in the set of regular facts and associated data and mechanisms than there is in the prose statement about men weighing more than women. If one defined the superiority of a set of statements as the amount of information present, then the set of regular facts is superior to the prose statement “on an average, men weigh more than women.” One can also learn facts about various classes of individuals, so one may have learned, as part of a definition of the class “men” and the class “women” that the average weight for members of the class “men” exceeds the average weight for members of the class “women.”

An informational fact, such as “Doug’s weight was 75 kilograms (using Scale 37, date March 23, 2013 at UTC 21:05:10, Doug was jumping up and down)” along with a set of other informational facts, would be superior to the regular facts, in that they provide more information than the regular facts. Assume that Scale 37 was used for all weights and actually displayed as the person’s weight their actual weight subtracted from 300 kilograms. Knowing this might allow one to conclude than the weights for the two sexes were such that women weigh more on an average than men.

With more information in regular facts than in raw prose, and more information in informational facts than in regular facts, one can reason to more possible outcomes and incorporate more conditions, when using IFacts. Informational facts more accurately reflect the real world than do other kinds of facts, and using informational facts leads to more conditioned outputs, what many might consider a more sophisticated or nuanced set of conclusions.

Authors often produce metadata for scientific datasets and authors can provide informational facts that represent the information in the article or prose itself. Datasets are used in the production of scientific articles and one can see in many expert systems informational facts derived from human experts. Note that few, if any, scientific datasets have anywhere near a full set of metainformation in the datasets.

Data, such as the mass for members of a species, could be translated into informational facts such as individual measurements using a scale. For example, person 1 might be weighed using Scale 1 and be determined to weigh 50 kilograms: <50 kilograms, Scale 1, person1>. One can attach more metainformation, and refer to a general weighing process, with the set of metainformation denoted within an IFact as

<50 kilograms, weighing with MI={Scale 1, July 23, 2013, UTC 15:42, 6 Hilltop Circle in Chapel Hill NC}, Person 1>.

An exhaustive description may be attempted, although one runs the risk of entering a large and decreasingly productive loop, where metainformation is needed to describe metainformation that is used to describe metainformation, and so on. A data dictionary can help address these points, by indicating the ranges of values that could be encoded in a smaller, finite way. Most informational facts will have incomplete metainformation. We assume that the more metainformation is present, the more accurately the informational fact reflects reality. There is a cost to reflecting reality accurately, and the value of metainformation needs to be considered, in combination with the cost.

Summary and Conclusions

The use of informational facts and associated metainformation, a form of metadata, will allow reasoning to be more accurate, and to minimize the compounding of errors when small errors are included in each observation. As more processing occur without human intervention, the error rate will increase; more accurate observations and metainformation about the observations will help minimize this increase in errors.

Facts may reflect the production of information in the world. When the input to these information producing processes, the nature of the process itself, and the information about the output of the process are gathered together, an informational fact is produced. Other forms of facts may be useful, but are often not based on empirical observations and may, or may not, be empirically shown to be true. The world may be described with a set of facts, and the possession of facts relevant to a particular issue may allow for the development of increasingly accurate deductions and inferences.

Metainformation associated with each informational fact describes the context in which the different parameters of the informational fact exist. Metadata about a fact, such as its topic, author, and so forth, are of great utility; these metadata are often produced by a later system. Metainformation often must be captured at the same time as the other parts of the informational fact. When the metainformation is captured at the time the fact is produced, the metainformation about the characteristics of a fact are more accurate than metainformation produced at some physical or temporal distance from the original production of the IFact.

Informational facts can be produced a number of different ways, and it is the metainformation that is associated with an IFact that makes them useful. Ideally, metainformation will be produced by the observer, person, or system who has direct access to the process and the objects that are the input to the process, as well as the informative output. When extracting IFacts from prose, for example, IFacts can be generated, possibly with flaws or omissions in the metainformation. Errors may occur with IFact extraction due to errors in the prose, errors in the extraction process, or other omissions. Author or observer generated IFacts and metainformation are the best way to produce accurate factual information for further analysis, by humans or computational systems, and will serve as the basis for “big data.”

Bibliography

Brassell, E. G. (2000). *Demonstrative Anaphora: Forms and Functions in Full Text Scientific Articles.* Chapel Hill, NC: Master's Paper, School of Information and Library Science, U. of North Carolina.

Foskett, A. C. (1996). *The Subject Approach to Information* (5 ed.). London: Library Association Publishing.

Garera, N., & Yarowsky, D. (2009). Structural, Transitive and Latent Models for Biographic Fact Extraction. *Proceedings of the 12th Conference of the European Chapter of the ACL*, (pp. 300-308). Athens, Greece.

Greenberg, J. (2005). Understanding Metadata and Metadata Schemes. *Cataloging and Classification Quarterly, 40*(3/4), 17-36.

Greenberg, J. (2009). Metadata and Digital Information. In *Encyclopedia of Library and Information Science* (pp. 3610-3623). New York: Marcel Dekker.

Karkaletsis, V., Fragkou, P., Petasis, G., & Iosif, E. (2011). Ontology Based Information Extraction From Text. *Multimedia Information Extraction. LNAI6050*, pp. 89-109. Berlin: Springer Verlag,.

Kastner, I., & Monz, C. (2009). Automatic Single-Document Key Fact Extraction From Newswire Articles. *Procedings of the 12th Conference of the European Chapter of the ACL*, (pp. 415-423). Athens, Greece.

Kneale, W., & Kneale, M. (1984). *The Development of Logic.* Oxford, U.K.: Clarendon Press.

Liu, X., Nie, Z., Yu, N., & Wen, J.-R. (2010). BioSnowball: Automated Population of Wikis. *KDD'10* (pp. 969-978). Washington, DC: ACM.

Losee, R. M. (2004). A Performance Model of the Length and Number of Subject Headings and Index Phrases. *Knowledge Organization, 31*(4), 245-251.

Losee, R. M. (2012). *Information From Processes: About the Nature of Information Creation, Use, and Representation.* Berlin: Springer Verlag.

Pasca, M. (2007). Organizing and Searching the World Wide Web of WWW 2007Facts - Step Two: Harnessing the Wisdom of the Crowds. *World Wide Web Conference 2007* (pp. 101-110). Banff, Alberta, Canada: ACM.

Siefkes, C., & Siniakov, P. (2005). An Overview and Classification of Adaptive Approaches to Informtion Extraction. *J. on Data Semantics, LNCS3730*, 172-212.

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Pacific Grove, CA: Brooks/Cole.

Weikum, G., & Theobald, M. (2010). From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. *PODS'10* (pp. 65-76). Indianapolis, IN: ACM.

Willis, C., & Losee, R. M. (2013). A Random Walk On An Ontology: Using Thesaurus Structure for Automatic Subject Indexing. *Journal of the American Society for Information Science and Technology*, In Press.

Winograd, T. (1972). *Understanding Natural Language.* New York: Academic Press.

Wittgenstein, L. (1981). *Tractatus Logico-Philosophicus.* London: Routledge.