A Vision of a Semantic Health Care Information Architecture[1]

Tom Munnecke,[2] David Booth, PhD[3]

In support of deliverable 3.1.4

August 29, 2013

Abstract

In December 2010, the President's Council of Advisors on Science and Technology (PCAST) issued Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward. This report made a number of recommendations, including the creation of a "*Universal Exchange Language that enables health IT data to be shared across institutions; and also to create the infrastructure that allows physicians and patients to assemble a patient's data across institutional boundaries, subject to strong, persistent, privacy safeguards and consistent with applicable patient privacy preferences.*" (p. 4)

The report also calls for a metadata oriented approach: *"The best way to manage and store data for advanced data-analytical techniques is to break data down into the smallest individual pieces that make sense to exchange or aggregate. These individual pieces are called "tagged data elements," because each unit of data is accompanied by a mandatory "metadata tag" that describes the attributes, provenance, and required security protections of the data…The indexing and retrieval of metadata-tagged data, across large numbers of geographically diverse locations, is an established, highly developed, technology—the basis of web search engines, for example...Innate, strong, privacy protection on all data, both at rest and in transit, with persistent patient-controlled privacy preferences, is likewise achievable, and must be designed in from the start."* (p. 4)

The report is critical of the Service Oriented Architecture (SOA) prevalent in much of today's health IT approach: *"In a sector as fragmented and rapidly evolving as*

[1] This paper was written with partial support from Army SBIR DHP12-004, Phase I deliverable 3.1.4

[2] Chief Technology Officer, The New Health Project

[3] Knowmed, Inc

healthcare, we believe it is impossible to build a national implementation of SOA solutions and directories that could be used and scaled indefinitely into the future. (To draw a loose analogy, the approach is like trying to enable free trade among hundreds of entities by negotiating a huge number of bilateral trade agreements). (p. 40)

The report lauds VistA: *"Some large healthcare organizations have overcome at least some of these barriers and successfully adopted EHR systems. The **VistA** system adopted by the Veterans Health Administration (VHA) has helped the Nation's largest integrated health system provide a highly regarded level of information technology supporting better care."*

What the report does not mention is that the initial architecture of VistA anticipated many of the features recommended in this report. VistA used a metadata-driven approach as the core of its functionality from its earliest implementation. The metadata approach, in a slightly modified format, was also the core of the DoD's Composite Health Care System architecture. The original VistA/CHCS (and Indian Health Service's RPMS) design ethos has many parallels to the recommendations of the PCAST report.

There is a thriving open source technology called Semantic Web or Linked Data, sponsored by the World Wide Web Consortium (W3c) that offers an existing technology that can meet the needs described for the Universal Exchange Language suggested by PCAST. (See Appendix A)

This technology provides a new "meta" level understanding of health IT. Just as algebra provides us with a higher level of understanding than arithmetic, this model provides us with a simpler, more powerful way of dealing with the complexities of health IT. And just like algebra, there is a learning curve associated with this new way of thinking that is not intuitive to those who are familiar with the lower level of abstraction such as provided by arithmetic.

Phase I of this project demonstrates a working prototype http://semantichealthcare.net/rambler showing how the VistA, CHCS, and IHS metadata dictionaries can be extended to a deployable version. This paper is the result of workshops in San Diego, MIT, and San Francisco, which helped refine the current topic, as well as build a network of experts which can serve as a foundation for future research into deploying the visions expressed in the PCAST report.

Recommendations for Stage II development

1. Stage 1 of this effort has demonstrated a prototype of a semantic approach to VistA and CHCS information at http://semantichealthcare.net This demonstrates on a limited scale the value of the PCAST Metadata approach. Deploying this prototype deeper into production use in Stage 2 will allow us to refine this model and get real-world experience at the interface level.
2. In order to make this a universal language that is capable of supporting the network effect as outlined in the PCAST report, a broader model needs to be

considered, embedding a Universal Health Language within a Universal Name Space that is an “umbrella” over today’s enterprise/interface model. Moving to a Linked Data model opens the door to scalable network effects, but also adds additional design responsibilities such as privacy, provenance, and other infrastructure. It also entails forming an open source ecosystem of users, clinicians, and companies to support this effort.

3. Extend the open source technical community supporting this technology.

Table of Contents

Introduction

The President's Council of Advisors on Science and Technology (PCAST) issued a report in December 2010 "Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward."[4] This report presented a bold vision of moving to a meta-data driven approach to solving the complexities of communication in health care. It advocated a framework that could take advantage of the "network effect," the dynamics that have fueled the web and Internet.

Time and again, over the past two decades we have seen innovative, entrepreneurial startups outside of health care transform how we interact with each other, how we do business, and how we get our news. Amazon, eBay, Craigslist, Wikipedia, Facebook, and Google are thriving because they offered a new path that was innovative – even at the expense of the traditional players who previously held powerful positions in their field.

Amazon did not try to integrate book publishers; it simply created a new online capability independent of them. Wikipedia did not try to integrate the encyclopedia powerhouses, it created a whole new model of "prosumers," both reading and writing the world's knowledge. Hierarchies of the past have crumbled in the face of network-centric approaches.

Health care is an enormously complex undertaking, and this complexity is growing at an astonishing rate with advances in the life sciences. We are discovering that social networks have an effect on our health – factors such as obesity, depression, and happiness are all affected by our network connections.[5] New technology allows us to track our physical activity, blood pressure, weight, insulin levels, sleep patterns, and more, all connected online. People have access to their own genomic information, and vast libraries of online medical information. Medical tourism, convenience care clinics in shopping venues, home health, and telehealth are also adding to the complexity of the landscape of medicine.

Adding to the intrinsic complexity of the process of medicine are the many political, economic, and legal issues facing our health care system. We have over 100,000

[4] "***Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward***."

[5] Cristakis, Nicholas A., and Fowler, Jame H: Caonnected: The Surprising Power of Our Social Networks and How They Shape Our Lives http://connectedthebook.com/index.html

pages of health care legislation, and the new ICD10 coding system introduces a 10 fold increase to 155,000 diagnostic codes.

Despite novel advances in technology, we are facing large-scale public health issues. Whooping cough, mumps, and measles cases are on the rise. The 2009 H1N1 Swine Flu epidemic infected about 60 million people, causing about 12,500 deaths. It is a question of when, not if, a more deadly, virulent flu appears – naturally or as an engineered pandemic.

The recent revelations over the NSA Prism surveillance system have focused the public's attention on privacy and security to an even greater extent than before. As we move to pervasive online health information, these concerns will be amplified.

Health care is becoming complex at an accelerating rate, and our will depend on our ability to adapt and cope with these rapid changes. Resilience and robustness of the whole system, not efficiency of disjointed subsystems, will be the key to success.

From one perspective, these factors are creating an impossibly complex situation. Exploding levels of complexity will lead to a complexity catastrophe that is far beyond any human or government level of understanding or control.

Another perspective is that using current tools and approaches will not facilitate a solution, but rather exacerbate the existing issues.

The PCAST report's advocacy of greater use of metadata provides the first step on the path to a solution. This uplifts our thinking to a higher level of abstraction – a "meta" level, which allows us to say less and less about more and more.

This is akin to moving from arithmetic to algebra when dealing with computational problems. For certain computations, arithmetic is a simple, effective tool. For more complicated problems, say calculating the area of a room, we can build tables to simply look up the length and width of the room to find the area without resorting to algebraic formulas.

Moving up the ladder of abstraction from "concrete" arithmetic thinking to "meta" level algebraic thinking gives us a much more powerful intellectual tool. Rather than seeing every problem as a new instance, we learn a *language* for dealing with computations. A single algebraic expression can relate to an arbitrary number of specific arithmetic problems. Having a language that allows us to think "Area = Length x Width" gives us the freedom to calculate any rectangular area, at any desired level of precision. The need for distinct tables for different scales and precision is an artifact of the limitations of arithmetic, which simply disappears with the higher-level algebraic formulation.

Currently it is as if we are trying to deal with our health care system using arithmetic, rather than algebra. We have decomposed our health care system into a massive collection of independent "arithmetic" level problems that are artifacts of our primitive way of looking at things. A "meta" level approach will give us a language to refer to these problems as simply specific instances of a more general formulation.

What appears impossibly complicated at the arithmetic level (i.e. our current health care system), becomes more manageable at the algebraic level.

The PCAST report's recommendations of the use of metadata and exploiting network effects provide a tantalizing glimpse at what this meta level looks like. Fortunately, there are a number of past experiences (the World Wide Web, Wikipedia, and VistA) to examine for lessons learned, and there are current technologies (Semantic Web, Linked Data, RDF) under development that might provide a path to a higher level of understanding.

The World Wide Web is arguably the most complex and globally transforming technology ever undertaken by mankind. Yet the initial technology underlying the web was amazingly simple. Sir Tim Berners-Lee, inventor of the web, said,

"What was often difficult for people to understand about the design of the web was that there was nothing else beyond URLs, HTTP, and HTML. There was no central computer "controlling" the web, no single network on which these protocols worked, not even an organization anywhere that "ran" the Web. The web was not a physical "thing" that existed in a certain "place." It was a "space" in which information could exist."[6]

Wikipedia is arguably the most comprehensive access point for world knowledge, used by 5 million people per month, with 130,000 active editors managing 31 million pages of information. Yet the original underlying technology – the Wiki – was written in a few afternoons by Ward Cunningham.

The VA's VistA system, one of the world's largest and longest running EHRs, originated as a very simple system, using a single language, one data type, 19 commands and 22 functions. [7]

A common critical success factor for the Web, Wikipedia, and VistA was the fact that they started simply, and became more complex through an evolutionary process between the designers and the users of the system. They grew through a process of adaptation by discovering what worked and then doing more of it. There is no "center" to the web, nor is there a "top" to Wikipedia. There is no hierarchy to the web, it is simply a vast space of objects to be viewed as the user wishes.

[6] Berners-Lee, Tim**, Weaving the Web, The Original Design and Ultimate Destiny of the World Wide Web**, Harper San Francisco, 1999, p. 36 and 209

[7] Video conversation between Ward Cunningham, inventor of the wiki, and Tom Munnecke, one of the original VistA software architects http://www.youtube.com/watch?v=pSnCN-HQXvI

Another common factor is that these systems were designed around a concept of a "space" rather than a collection of interfaced systems. The web did not try to interface AOL and Prodigy, nor did Wikipedia try to integrate World Book and Encyclopedia Britannica, nor did VistA try to "integrate" lab and pharmacy systems. Interconnectivity was simply an intrinsic property of the information space. A web user can drag a book's URL from Amazon.com to Twitter without an interoperability agreement, committees, or an Application Programming Interface (API). It is simply something that web browsers do.

These successes contradict the notion that large, complex entities require large, complex management efforts to manage them. Indeed, the fact that they did not have large complex management was crucial to their success.

Analysis of The PCAST Recommendations

Move to Linguistic Expressions, rather than APIs

The PCAST report's recommendation to use metadata is a call to use language and linguistic expressions to communicate. This is much more powerful and flexible than using "hard coded" APIs (Application Program Interfaces) that provide only a function in the context of a program.

We see this linguistic approach in the design of the web. The web is driven by linguistic references, not APIs. For example, a web browser interpreting an HTML page can follow a link to a photograph with a .jpg type or a document with a .pdf type. There are not separate APIs for the different document types, just different names. Internally, the browser may interpret the pages with references to APIs, but the web itself runs off names.

Information Space, rather than Interfaced Enterprises

One of the original visions for interfacing the VA (VistA) and DoD (CHCS) systems was to create a universal namespace, linking the data dictionaries of the VA and DoD, and, by extension, any other entity wishing to link data. This could have been implemented as metadata linking the metadata at each site, or as a single, "flat" meta-level description of all the information. Today's Semantic Web technology provides a tool for building this "space" of linked information, independent of the particular hierarchies that may hold the information. The Project's website http://semantichealthcare.net illustrates a way of browsing the existing metadata of VistA. One of the critical first steps is to adopt a Universal Resource Identifier for a national health information space.

Improved Privacy and Security

This information space would be based on a fine-grained object level, so that access to the information could be mediated at whatever level is appropriate for the specific information and interaction.

Network Model, rather than Hierarchy

A Semantic Web model employs a network of interaction, not constrained to predefined hierarchies. This is a little like finding things on Google according to whatever search terms you like, rather than using a Card Catalog, trying to find information according to the hierarchies predefined by the Dewey Decimal System. If information is truly hierarchical, (for example, a family tree), hierarchies can be superimposed on the network.

Meta Level Design

Once we have a meta language for describing the operation of our information space, then we can also move our design thinking to this meta level. This allows us to think at an "algebraic" level of abstraction, rather than repetitive "arithmetic" exercises. The Meta Level design approach allows us to "dissolve" problems that previously required independent solutions.

Integration Crunch

The core of the VistA/CHCS architecture was that the system would begin with a common set of "umbrella" tools and databases, which would then be specialized to specific functions. All applications would use the same active meta data tools, and have access to the common patient database. If one application wanted to manage pharmacy inventory and another wanted to manage other medical supplies, they would begin with nearly identical logic, applied to different files. The pharmacy would require DEA information for certain items, and the medical supply inventory would require FDA information. These differences could be handled at the meta level – in the data dictionary definitions of the information being processed. VistA started with the whole and then differentiated into parts. The goal of VistA was to create a path of least resistance to a cohesive, integrated system. VistA was integrated by virtue of its not having disintegrated.

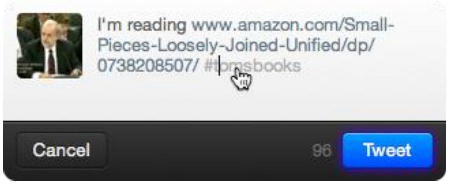
An alternative approach has been to start with the pieces and then integrate them. At smaller scale, this may be a reasonable approach. A pharmacy inventory system could be interfaced to a medical supply inventory with some effort. But as the number of parts increases, it becomes ever-more complex to integrate them into a workable whole. What looks simple at small scale becomes explosively complex at larger scale.

The way out of this integration crunch is to move to a set of tools that works at this meta level. The meta data approach of VistA/CHCS fits nicely with the metadata recommendations of the PCAST report. The experience in scaling up VistA and CHCS gives us a foundation for understanding the issues of large scale deployment.

On the Difference between Exchanging and Linking Information

People can share information on the web simply by dragging a URL from one web site to another. This example will illustrate some of the tools that allow this model of linking, rather than exchanging information.

Imagine that someone wanted to share information on a book found on Amazon.com. The person could copy the URL of the book's description on Amazon, the send it as a tweet through a Twitter account:



I'm reading www.amazon.com/Small-Pieces-Loosely-Joined-Unified/dp/0738208507/ #tomsbooks

As part of the tweet, comments can be included ("I'm reading") along with the URL of the book on Amazon (http://www.amazon.com/Small-Pieces-Loosely-Joined-Unified/dp/0738208507/), and with some additional *metadata* (hashtag #tomsbooks).

This is an example of *linking* a page on Amazon to a twitter feed and to the *#tomsbooks* hash tag set for others to search.

There was no *exchange* of any information about the book. If someone reads the tweet, they are able to click on the link to Amazon and read more about it. They can do this from a smart phone, a computer, or any web browser, anywhere in the world. If a new technology comes along, it would be able to follow these links as well.

This illustrates several interesting things about the web:

1. The URL of a web page is a universal identifier. This is a founding principle of the web. It links a name to an information object, independent of where, when, or how the reader uses it. Contrast this with the traditional telephone exchange model. If there was an exchange of phone numbers, for example, the actual sequence dialed might vary depending on whether it was dialed it from a mobile phone or a landline, from within a business or not (dial 9 to get out), whether the caller was in a 10-digit area code zone, inside the area code or not, or inside the country or not. If someone move to another area code, their landline number changes, and everyone else dialing them has to change their address books.

2. Neither Twitter nor Amazon had to interface with each other. Sharing the information simply required a reference to the book on Amazon (the URL contained the information Amazon needed to figure out the book being referenced) that was pasted into a Twitter message. The web standards (URL, HTTP, HTML) provided the "glue" to make this all work. Linking to Barnes and Noble, a blog, a Facebook page, or whatever else the user selected would have worked with the same results.

3. Using an API (Application Program Interface) was unnecessary to link Twitter to Amazon. The linking of information was done automatically via the web protocols.

4. It was very simple. All that was required was copying the URL from one page and pasting it into another page. These requests triggered off many complex computer

interactions, perhaps involving thousands of computers in hundreds of locations, but to the individual sharing the information, it was a simple copy-and-paste activity.

5. The person sharing the information did not need to test it to be confident that it would work. Even if Twitter was entirely new, the individual would have known that a URL would work anywhere. If someone got the URL, they could access the information. If desired, it would be simple to test the procedure by tweeting the message, then send an email to someone to see if they could read the tweet.

6. The web is scalable. Any number of websites can participate in the web. The more sites there are, the more potential there is for linking, which makes the web more valuable for all concerned. This *network effect*, mentioned in the PCAST report, is a powerful tool for driving change. The information space of the web does not get used up like physical space in a shopping center gets used up with stores. The web, exploiting the network effect, moved us to a notion of network abundance, rather than scarcity. The problem on the web is not a lack of information, but rather the means to make sense of all the information.

7. The URL of the book was treated as a publicly accessible piece of knowledge. Amazon wants people to link to them, see their book offerings, and download or purchase their books. The visibility of a page, however, is a property of the information itself. For example, someone can publish an Amazon wish list publicly:

Based on an individual's preferences, a wish list can be made public so as many people as possible can view it. However, there may be other wish lists that the individual wants to keep private, perhaps to friends and family, or even for personal use only. When a wish list is setup with Amazon, that option is available.

The Amazon pages relating to credit card information, however, are controlled by a logon process. Even if someone tweeted an account management page, the URL in people's browser is inaccessible unless they can log on to the account through Amazon's security procedures.

The information that is on Amazon thus ranges from completely open to material that is individually controlled, to things that are controlled by Amazon's security system, to information that is Amazon-internal, and not even viewable to outside users via a browser.

The Importance of Language

One of the most profound implications of the PCAST report is the recommendation to move to a linguistic framework for dealing with health information. Language has a profound influence on the way we think. Crafting a language also crafts our patterns of thought, and also defines the limits about what we can think about.

Unix Programmers are familiar with a command shell from which to control their computers. Despite the proliferation of types of shells, this is one of the unifying aspects of the Unix community. A language creates a linguistic shell within which its speech community can communicate. However, language cannot describe what it cannot talk about. For example, a language of disease cannot talk about a language of health.

The Unix shell is a very powerful set of tools that can control just about everything happening in the computer. But imagine if Unix had been designed only to sort data, and the shell only had terms for naming files, ways of sorting them, and where to put the results. This “Sort Shell” would not be able to describe the fact that it could not schedule jobs - it could not reach out of its linguistic shell to describe what it cannot do. The Unix shell is a broader, universal linguistic shell that can talk about sorting, scheduling, network distribution of jobs, etc.

The lesson learned here is that if we are to have a “universal” language of health, it must be a broad “umbrella” language. If we start with a language of disease, we cannot pop out of it to talk about health. If we start with a language based on causality using on Gaussian distributions, we cannot pop out of it to talk about cascades based on power-law distributions. If we start with a finite-state grammar, we cannot pop out to use generative grammar. If we have solely a finite state language of disease and symptoms, we cannot pop out of it to understand homeostasis, mind/body interaction, the generative language of the immune system, (social) network effects of health, our evolving understanding of genomics and the life sciences, and many other aspects of health.

“Popping out” of a language can be a very difficult thing to deal with, particularly if the language is tied to political, economic, or professional identity. For example, the American Psychiatric Association makes a majority of their revenue creating the DSM language. Language is often used to isolate and confirm professional jargon, not necessarily enhance communication.

Missing Nothings

Imagine that someone from today was transported back to ancient Greece and tried to communicate what we know today about mathematics. They would be using the roman numeral system for counting: I, II, V, X, etc. Most people would be able to count using this system.

However, things would get complicated when we tried to explain our decimal numbering system. “Why get so complicated?” they would ask. “We can count everything we can see, and we can see everything we can count.” We would say, “Well, there are things you can do with decimal numbers, ways of calculating, far more powerful than you can imagine with your numeral system.” The notion that we could use math to control rockets flying to the moon, calculate their mass and acceleration as they produced thrust, burned off fuel, zoomed around in orbit, would be unbelievably complex to them.

The deal breaker, though, for the excursion would likely have been the concept of zero. The Greeks had no notion of zero. They started with one and counted upward. They had no place holders the way we have 10s, the 100s, etc. They just had an ever-increasing string of numbers.

The very idea that we would have a symbol to represent “nothing” would have been inexplicable. Why have a formal representation for something that does not exist? Why have such a complex collection of digits, place holders, and zero when we have a perfectly complete “one tally mark, one item” system that everyone can understand?

The Roman numeral system was concrete, and simple to understand. The decimal system was abstract, so abstract that the concepts of algebra or calculus would have been beyond the ancient Greeks’ wildest imaginations.

Despite the ancient Greeks’ sophistication in other areas of mathematics, their ignorance of zero preventing them from moving up the ladders of abstraction that are familiar to even the youngest school child today. Zero was a *missing nothing* to them.

It was not until the time of Isaac Newton and Gottfried Wilhelm von Leibniz that the power of zero was fully demonstrated, through the Calculus of the Infinitesimal. Basically, what happens to things when we look at numbers as they go to zero? Algebra consists of manipulating an equation by adding nothing to it until it is solved.

The “meta” level of algebra was invisible to the ancient Greeks, because zero was a blind spot, something outside their linguistic shell. Zero was a “missing nothing” to them, whose absence blocked their intellectual growth into algebra and calculus.

This brings up an intriguing question: are we currently being blocked by a “missing nothing” in the linguistic shells we have today? Is there a blind spot in our current thinking that precludes us from understanding the whole picture of health? Would this make things that appear to be imponderably complex simple again? We will never know until we discover and develop a language can address these issues.

Connections and Dots

One way of looking at complex systems is to think of them as having all the “things” in the system represented by dots, and all of their interrelationships represented by lines showing the connections.

In the pre-Google days of the library, the books would have been the dots, and the connectors would have been the footnotes, references, card catalog cross references, and the like.

The Dewey Decimal system was one of the more popular ways to corral the dots into meaningful clusters. Mr. Dewey categorized #393 at *Death Customs*, #435 as

German Grammar, and #539 as *Modern Physics*, #635 as Shorthand, and #376 is no longer used, but used to be *Education of Women.* The whole field of *Modern Physics* was given as significant of a place on the abstraction hierarchy as *Death Customs.* And antiquated terms such as *Education of Women* still bounce around the card catalogs and bookshelves. We will call this the pigeonhole paradigm, seeking to create a predefined structure that has a place for every datum, and every datum in its place.

Tim Berners-Lee did not try to apply this approach to the web. The web's information space is not just a collection of dots, but also the connectors between the dots. Computers can look at the connectors as well as the dots, and use their own ways of finding things.

Beyond the immediacy of access to today's information, the space of the web (represented through the ease of a semantic Google search) represents the dissolution of hierarchy. An individual looking for information would not see the Dewey Decimal System, librarians announcing closing time, or having to wait for interlibrary loans to exchange information.

Rather, the current information space is far removed from the era of going to a card catalog at the library, looking things up with the Dewey Decimal System, and asking for interlibrary loans. There is a connected world; a world where things are networked into a fabric of people, smart things, and agents.

When Tim Berners-Lee invented the web, he did not try to organize it. He knew he was creating a chaotic mess of URLs. He did not try to assign web sites 1-100 to physics, 101-200 to chemistry, or try to adopt some kind of Dewey Decimal System to overlay a predefined hierarchical coding system over the web.

Nor did he try to invent Google. Many search engines came and went, seeking to crawl the web and provide a way to discover information based on content, not structure.

Yahoo tried to do a hierarchical index, ("Yet Another Hierarchical Organized Order). At one time, it hired a team of employees seeking to organize the web, into a "card catalog" approach. This effort eventually collapsed, leaving Yahoo at the mercy of other search engine companies.

One of the reasons that Yahoo's "curated index" approach to the web collapsed is that it did not scale. It was simply not possible to hire enough people to keep track of all the information on the web.

Google, on the other hand, was designed around the notion of scale from the outset. He and Larry Page sought to push the limits of how far they could build a scalable collection of commodity computers, and search was a vehicle for testing out their ideas.

Google does not publish the “secret sauce” of it search strategy, but it probably uses a variant of some form of Latent Semantic Analysis, creating a very high dimensional space to relate terms to each other. In this space of perhaps hundreds of thousands of dimensions, “foot covering,” “shoe,” and “boot” cluster together. Search users do not know that that this is happening, nor that their selection choices help refine future searches. Google exploits the network effect[8] of growth. The more people use Google, the smarter it gets.

Tim Berners-Lee allowed the web to be broken. Previous hypertext efforts preceding his tried to insure bi-directional referential integrity. If A pointed to B, B had to point back to A. Relaxing this notion of perfection, he allowed the infamous “404 Not Found” error. While this may be vexing at times, it is necessary feature to make the web resilient and adaptive to changing needs.

He did not try to “integrate” the networks of the time. AOL, Prodigy, and others all had proprietary networks based on connection time. Users paid a fee for connection time. If someone wanted to look up something on AOL and something else on Prodigy, they needed two accounts, logging on and off between them, paying the connect charges. This is a “castle and drawbridge” architectural model. Information is kept inside the castle, and gatekeepers seek to ensure privacy, control, and information flow by watching their drawbridges. Only trustworthy visitors would be allowed across the drawbridge.

Similarly, Pierre Omidyar did not try to integrate auction houses when he created eBay. Craig Newmark did not try to integrate newspapers classified ads sections when he created Craigslist. Jimmy Wales did not try to integrate World Book and Encyclopedia Britannica when he invented Wikipedia. Instead, they each created a novel “space” for their information to appear, independent of the powerful hierarchies of the powers-that-were before their successes.

Today’s call for increased health IT interoperability, standardization, predefined hierarchical coding systems, and centralization flies in the face of the success we have seen in other sectors over the past decades. A web user can simply drag a URL from an Amazon search to his Twitter feed. There is no need for an interoperability agreement between Amazon and Twitter. There is no need for standardized book-to-short-message exchange formats, nor an API to accomplish it. The user could find a book via Google search, Amazon’s recommender system (readers who bought this also bought ….), the “like” of a Facebook friend, a reference in a paper, or any other

[8] The network effect can be illustrated with the growth of the fax machine: the first fax machine was worthless. It only had value when the second fax machine appeared. Each succeeding fax machine created value for all the other machines. The fact that this example current during the dot com boom of the 1990’s, is dated says much about the pace of technology. The underlying network effect is still very much current.

approach. There is no need for a Dewey Decimal System around which to organize books. There is no need for centralization: the reader could have just as easily used a Barnes and Noble site, a used book seller, a print-on-demand provider, or a local library.

The URL provides a name that is universal – it is the same whether you are looking at it in a Tweet, a web page, or an email. It is the same whether you use it over a wireless phone, a wired PC, or a tablet on an airplane WiFi.

Conclusions

The PCAST report recommendations are a viable innovation towards improving health IT, which introduce a fundamentally different way of looking at information – a network approach in contrast to traditional hierarchies.

Deploying a Semantic Web based model has potential to significantly improve VA/DoD interoperability and communication, as well as prepare for future advances in life sciences, web technology, and health care.

It has potential for opening up health IT to the network effect that has driven much of the web's revolutionary capabilities, but this still requires exploration and development. Moving to a "meta" level language and operation has tremendous potential to simplify health IT.

This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. **W81XWH-13-C-0042**.

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

Glossary

Linked Data	To make the Web of Data a reality, it is important to have the huge amount of data on the Web available in a standard format, reachable and manageable by Semantic Web tools. Furthermore, not only does the Semantic Web need access to data, but *relationships among data* should be made available, as well, to create a *Web* of Data (as opposed to a sheer collection of datasets). This collection of interrelated

	datasets on the Web can also be referred to as Linked Data. http://www.w3.org/standards/semanticweb/data.html
Metadata	Metadata is machine understandable information about web resources or other things (Per Tim Berners-Lee's 1997 discussion) http://www.w3.org/DesignIssues/Metadata
OWL	The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. OWL is a computational logic-based language such that knowledge expressed in OWL can be reasoned with by computer programs either to verify the consistency of that knowledge or to make implicit knowledge explicit. http://www.w3.org/2001/sw/wiki/OWL
PCAST Report	President's Council of Advisors for Science and Technology 2010 White Paper, *"Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward."*
RDF	Resource Description Format RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. http://www.w3.org/RDF/
Semantic Web	The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS. http://www.w3.org/standards/semanticweb/
SKOS	http://www.w3.org/2001/sw/wiki/SKOS Simple Knowledge Organization System is a common data model for sharing and linking knowledge organization systems via the Web. Many knowledge organization systems, such as thesauri, taxonomies,

	classification schemes and subject heading systems, share a similar structure, and are used in similar applications. SKOS captures much of this similarity and makes it explicit, to enable data and technology sharing across diverse applications. The SKOS data model provides a standard, low-cost migration path for porting existing knowledge organization systems to the Semantic Web. SKOS also provides a lightweight, intuitive language for developing and sharing new knowledge organization systems. It may be used on its own, or in combination with formal knowledge representation languages such as the Web Ontology language (OWL).
SPARQL	SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports aggregation, subqueries, negation, creating values by expressions, extensible value testing, and constraining queries by source RDF graph. The results of SPARQL queries can be result sets or RDF graphs. http://www.w3.org/TR/sparql11-query/
Universal Health Exchange Language	A language called for in the President's Council of Advisors for Science and Technology 2010 White Paper, *"Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward."* "PCAST has also concluded that to achieve these objectives it is crucial that the Federal Government facilitate the nationwide adoption of a universal exchange language for healthcare information and a digital infrastructure for locating patient records while strictly ensuring patient privacy." This is the language of the Universal Health Space Framework, named the Universal Health Language. http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf
URI	Universal Resource Identifier: a unique identifier, globally accessible across the entire Universal Health Space, subject to constraints of privacy and security.

Workshops and Collaborators

Over the years, Tom Munnecke has developed a workshop format for exploring significant issues. Each workshop is in a salon format with a small group of carefully chosen "grounded eclectics" – people who are experienced in some specialty, but are also capable of lateral thinking. Workshops may have computer scientists, doctors, economists, science fiction writers, anthropologists, and others in a "slow conversation" format. Workshops deal with topics that are large and complex, the goal of the workshop is not to solve a specific problem at one sitting, but rather to build an ongoing network and conversation around the topic.

We held three workshops (two in San Diego, one at MIT) and attended one conference (in San Francisco) in conjunction with Phase I of this project:

Science of the Individual	Nov 14, 2013 Encintas, CA	http://www.new-health-project.net/2013/03/21/workshop-report-science-of-individual/
Semantic Health Workshop	April 19, 2013 MIT (W3C offices)	http://www.new-health-project.net/2013/04/23/report-on-semantic-health-workshop-at-mit-april-19-20-2013/
RDF as Universal Health Language (conference)	June 3, 2013 San Francisco	http://www.youtube.com/playlist?list=PLt-A972QBADUZZK9tIzNhDSPb6gv-Uvog
RDF as Universal Health Language (workshop)	June 25-26, 2013 Encinitas, CA	http://www.new-health-project.net/2013/06/26/report-on-workshop-rdf-as-a-universal-health-language-encinitas-june-25-26-2013/

Collaborators

This is a list of some of the participants at these workshops. We wish to acknowledge the participation of the many people who participated, and the many contributions they have made. It is impossible to include all the ideas in a single paper, but most of the presentations are freely available as video on the web.

The author wishes to thank the many contributors for the input, but the final product of this effort is the author's responsibility, who assumes responsibility for for any errors or possible misinterpretations of the work.

Adrian Gropper, MD, CTO, Patient Privacy Foundation
Alex Tan, Frog Design
Brian Ahier, www.ahier.net
Christophe Lambert, PhD, CEO, Golden Helix
Conor Dowling, CTO, CareGraf video
David Booth, PhD, KnowMed
David Brin, PhD, Science Fiction writer/futurist, author of The Transparent Society,

David Kronenfeld, PhD, Professor of Anthropology, UC Riverside
Eric Prud'hommeaux, W3C Life Sciences group
Eric Von Schweber, Surveyor Health
Gio Wiederhold, PhD, Professor of Computer Science, Stanford University, video
Harold Koenig, MD, Vice Adm (ret), former Surgeon General of the Navy video
Heather Wood Ion, Health Care Consultant video
Joanne Luciano, PhD, W3C Semantic Web Life Sciences group
John Mattison, MD, CMIO, Kaiser Permanente
Linda Von Schweber, Surveyor Health
Luis Alvarez, PhD Software Specialist, Kitware
Nancy Tomich, Managing Director, New Health Project
Peter Norvig, PhD, Director of Research, Google, video
Thomas Payne, PhD, Professor of Mathematics, UC Riverside
Vanessa Moeder, genomics consultant
Vernor Vinge, PhD, Science Fiction writer/futurist, invented the term “Singularity” video
Wesley Turner, PhD, Software Specialist, Kitware

Appendix A: RDF as a Universal Healthcare Exchange Language: Realistically Achieving Semantic Interoperability

David Booth, Ph.D.
KnowMED, Inc.

Latest version of this document: http://dbooth.org/2013/rdf-as-univ/rdf-as-univ.pdf
Slides: http://dbooth.org/2013/rdf-as-univ/slides.pdf

Abstract

Electronic healthcare information is currently represented in a bewildering variety of incompatible data formats, models, and vocabularies. To improve healthcare effectiveness and efficiency, healthcare computer systems need to be able to exchange electronic healthcare information in a machine processable form that enables semantic interoperability between systems. This paper explains how a *universal healthcare exchange language* can be realistically adopted in order to achieve semantic interoperability between healthcare systems. It explains why RDF is the best available candidate for such a language, and it outlines research needed to prove viability on a national or international scale.

Introduction

Imagine a world *in which all healthcare systems speak the same language with the same meanings covering all healthcare.* True semantic interoperability: what would it be like?

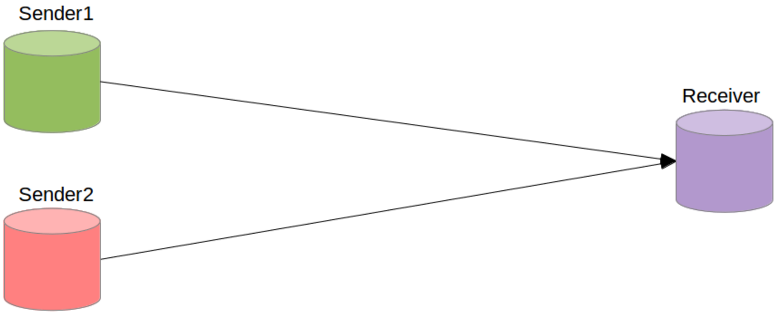
- *Better treatment*, as doctors could more easily obtain an automatically integrated view of a patient's condition and history.

- *Better research*, as researchers could more easily combine and analyze data from many sources.
- *Lower cost*, as efficiency would be improved.

Unfortunately, electronic heathcare information systems today are a Tower of Babel, using hundreds of different and incompatible data formats, models, and vocabularies, thus inhibiting semantic interoperability. The President's Council of Advisors on Science and Technology (PCAST) highlighted the importance of this problem and called for a *universal exchange language*: "PCAST has also concluded that to achieve these objectives it is crucial that the Federal Government facilitate the nationwide adoption of a universal exchange language for healthcare information".[1]

What is semantic interoperability?

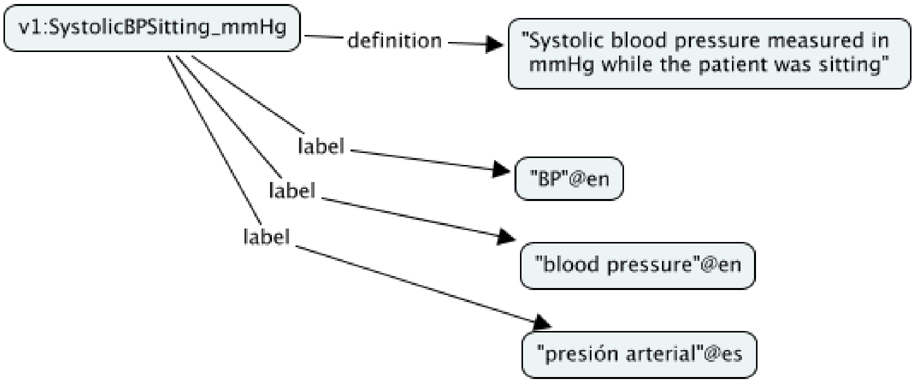
Suppose a patient is being treated by a healthcare provider (the information Receiver), and that patient's medical records are requested from two other healthcare providers (Sender1 and Sender2) in order to automatically obtain an integrated view of the patient's condition and history. If semantic interoperability were achieved, the Receiver system, given appropriate authorization and routing information, could: (a) obtain the patient's medical records from Sender1 and Sender2; (b) combine those records into an integrated patient record; and (c) extract any needed information from the result -- all automatically.



To achieve this automatically (without human labor), the information must be expressed in a **machine processable** format that computer systems can properly interpret, both efficiently and reliably. A machine processable format is a structured format that enables a computer system to properly "understand" the data, i.e., to make meaningful use of it. This is in contrast with human-oriented information formats, such as narrative text or scanned documents. Human-oriented information formats can be efficiently and reliably generated from machine processable formats, but not vice versa.

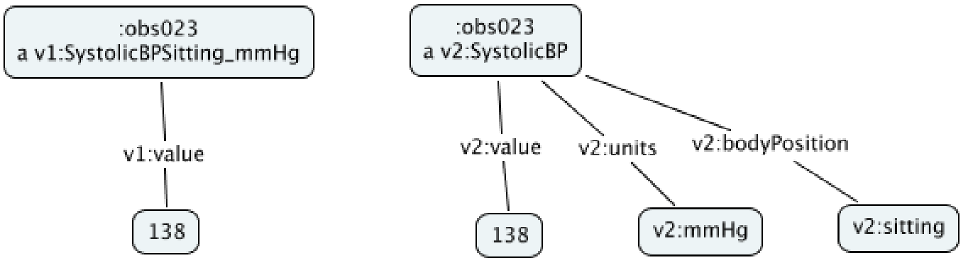
Even if a data format is understood, for the Receiver system to properly interpret the received information it contains, the information must also be expressed in a **controlled vocabulary** that the Receiver understands. However, the use of standard medical terms is not enough to ensure accurate interpretation, because the same term may be understood differently by different parties -- sometimes in subtly different ways. For example, to assess a patient's health risk factors, a data element may capture information about whether the patient smokes. But the term "current smoker" may have different meanings in different contexts[2]. To ensure accurate interpretation of captured data, terms in a controlled vocabulary must have agreed-upon *definitions*.

On the other hand, sometimes different natural language terms mean the same thing, i.e., they have the same definition. For example, "blood pressure", "BP" and "presión arterial" (Spanish) may all refer to the same concept of blood pressure. Thus, to facilitate machine processing, avoid ambiguity, and enable meaningful information display to human users, it is helpful to represent each term in a controlled vocabulary as an **unambiguous concept** consisting of: a unique *identifier*; a *definition*; and one or more human-oriented *display labels*. For example if v1 is a controlled vocabulary then v1:SystolicBPSitting_mmHg may be a globally unique identifier for a particular concept of a blood pressure measurement, having an associated definition and multiple display labels.


v1:SystolicBPSitting_mmHg
definition
"Systolic blood pressure measured in mmHg while the patient was sitting"
label
"BP"@en
label
"blood pressure"@en
label
"presión arterial"@es

The Receiver system does not necessarily need to directly understand each controlled vocabulary term it encounters, but if it encounters a term that it does not understand, then there should be a standard, automatable algorithm to enable the Receiver to obtain an accurate definition of the term. Whenever possible, the definition should be provided in both a machine-processable form and a human-oriented form that relates the term to other terms that are already known to the system, so that the system can bootstrap its understanding. For example, in **Linked Data**[3], each term is unambiguously identified by a URI[4]. If a system does not already understand a term, the system can *dereference* the term's URI to obtain the term's definition.

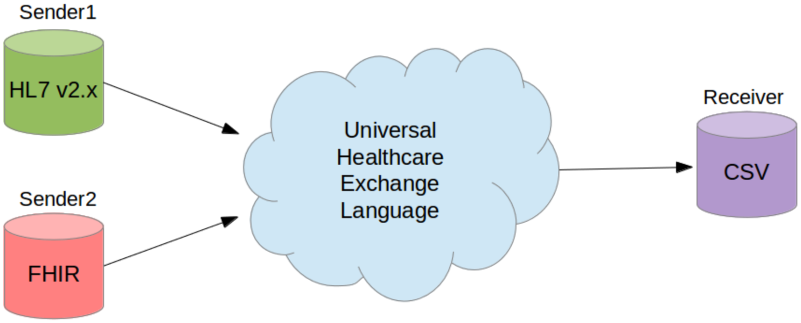
Another complexity in semantic interoperability is that different parties may represent information using **different data models**. For example, a systolic blood pressure measurement might be encoded either in a "pre-coordinated" style that also indicates the body position and the units of measure, or in a "post-coordinated" style that breaks the information into atomic pieces, as shown below. [Thanks to Stanley Huff for this example.] Both representations carry the same information.



Fortunately, **metadata** concepts in a vocabulary can **signal the expected data model**. Such metadata should be included with the instance data to enable a receiving system to properly interpret the data. For example, if an observation is recorded as a v1:SystolicBPSitting_mmHG measurement (using vocabulary v1), then the units and body position may already be implied by that concept's definition. Alternatively, if the observation is recorded as a v2:SystolicBP measurement (using vocabulary v2), then the use of that identifier (v2:SystolicBP) signals the use of a data model in which the value, units and body position are explicitly indicated as associated data elements, using terms v2:value, v2:units and v2:bodyPosition. In summary, the vocabulary defines a set of concepts along with their associated data models, and the data models indicate the relationships between the concepts. Such a vocabulary is sometimes called an *ontology*. Finally, because a concept can signal what data model is used, the problem of standardizing the data models boils down to the problem of standardizing the concepts: if the concepts are standardized, the data models implicitly become standardized.

The role of a universal healthcare exchange language

In theory, semantic interoperability could be achieved by converting all healthcare systems to internally use one standard format and vocabulary. However, this option is neither politically feasible nor advisable, both because of the enormous transition investment that it would require and because it would stymie innovation. Hence **we will assume that healthcare systems will continue to internally use whatever data formats and vocabularies they choose**, and examine what is needed to achieve semantic interoperability in the *exchange* of healthcare information between systems. To illustrate, let us assume that Sender1 uses one format (HL7 v2.x[5]), Sender2 uses another (FHIR[6]), and Receiver uses a third (CSV, comma-separated-values[7]).



To achieve semantic interoperability, the data from Sender1 and Sender2 must be somehow transformed into the format and vocabulary that the Receiver can understand. The purpose of a *universal healthcare exchange language* is to simplify this transformation by enabling healthcare information to be transformed to and from a common intermediate language. Although the use of an intermediate language means that two transformations must be performed during each information exchange instead of one – source to intermediate language, and intermediate language to destination – it dramatically reduces the implementation complexity by reducing the number of kinds of transformation that are needed. Instead of needing *(n-1)*(n-1)* transformations only *n* are needed, where *n* is the number of distinct kinds of sender/receiver languages. If healthcare systems eventually choose to use this common language internally, all the better, but that is not a requirement for achieving semantic interoperability.

Why is semantic interoperability so difficult?

If a universal healthcare exchange language can be standardized, and all parties exchange healthcare information using the same standard data models and vocabularies, then semantic interoperability can be achieved. This sounds simple, but it is deceptive. It would be easy enough to standardize a sufficiently flexible *syntactic* framework for such a language. But standardizing the *semantics* -- the data meaning -- is far more difficult. Data meaning is determined by the data models and vocabularies that are used in the data.

There are three main reasons why these are so hard to standardize:

- Medicine is very complex, involving many thousands of interrelated concepts and many overlapping areas of expertise.
- As the size and complexity of a standardization task grows -- and the number of committee members grows -- the rate of progress diminishes toward zero.
- Medical science and technology are continually changing, requiring new concepts all the time. It is a moving target.

In short, it is not feasible to stop the world until we can standardize all of the data models and vocabularies needed for a universal healthcare exchange language to achieve full semantic interoperability. Instead, we need an approach that acknowledges and accepts the dynamic nature of the problem.

Key requirements for a universal healthcare exchange language

Because of the above challenges, a viable universal healthcare exchange language must meet several key requirements:

- The language must accommodate the **continual incorporation of new or revised data models and vocabularies**.
- The language must support the use of **existing and future healthcare information standards**, to reap the benefits of standardization whenever possible.
- The language must support **decentralized innovation**, to avoid committee bottlenecks and enable new concepts to be used before they have been standardized.

Furthermore, to support the graceful adoption of new data models and vocabularies:

- The language should enable new or revised data models and vocabularies to be **semantically linked** to existing data models and vocabularies.

- The language should enable **authoritative definitions** of new concepts to be obtained automatically, so that when a system encounters a new term, the system can properly understand it.

The *best available candidate* to meet these requirements is RDF.

What is RDF?

Resource Description Framework (RDF)[8] is an information representation language. An international standard, RDF is qualitatively different from other information representation languages currently common in healthcare, due to its *schema promiscuity* (explained below), its basis in web standards, and its emphasis on semantics. RDF acts as a unifying substrate for healthcare information, allowing any number of vocabularies and data models to be semantically connected and used together. How this works, and why this is important, will be explained in the remaining sections. RDF does not and cannot solve the semantic interoperability problem by itself -- no technology can, as there are important social factors at play in addition to technical factors -- but RDF can be a crucial component as a universal healthcare exchange language.

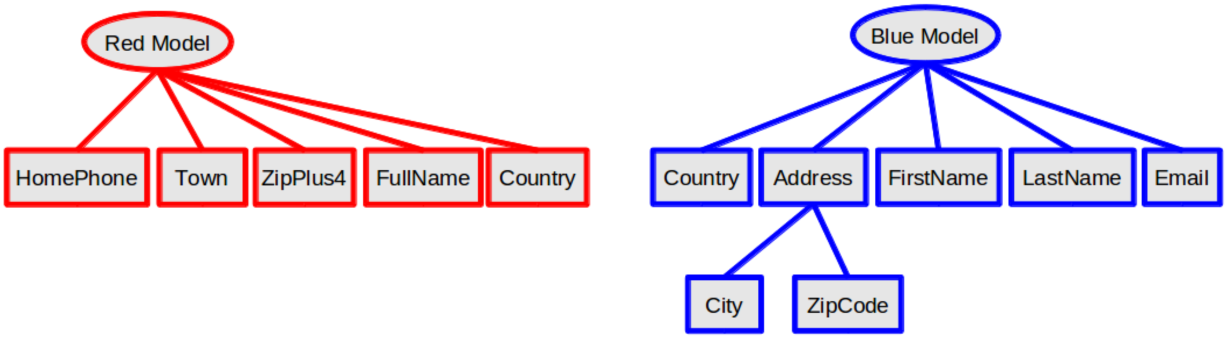
URIs as unique identifiers

URIs are the basis of the World Wide Web, and can be used both as web page locators and as globally unique identifiers. RDF uses URIs as globally unique identifiers. Any concept used in healthcare can be given a URI as an identifier, including concepts in vocabularies, procedures, medications, conditions, diagnoses, people, organizations, etc. The use of URIs enables orderly decentralized allocation of identifiers, and, if Linked Data principles are adopted, it means that an identifier can be conveniently associated with its definition: a URI can be dereferenced to a web page to find its definition.

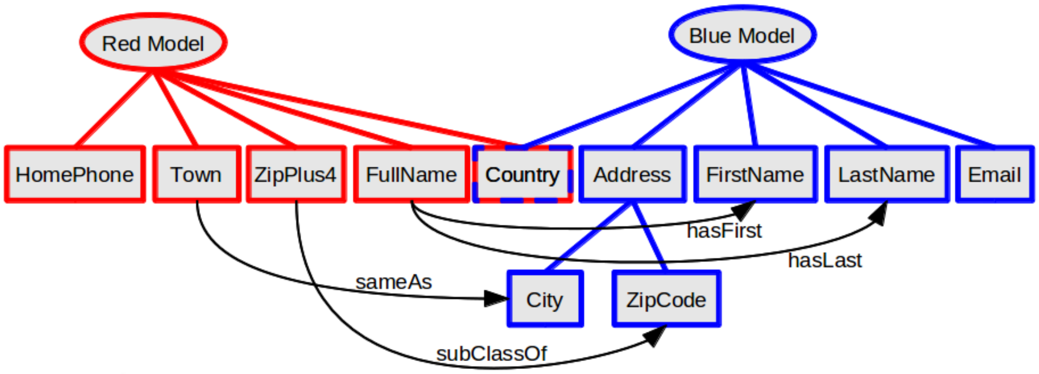
Schema promiscuity

The most important characteristic of RDF that makes it well suited to be a universal healthcare exchange language is what we may call *schema promiscuity*. In contrast with most common information representation languages such as XML, RDF allows data expressed in different data models or *schemas* to peacefully coexist within the same datasets, semantically interlinked. Relationships between concepts in different models can be expressed in RDF just as relationships within a model are expressed. Instead of requiring one model to incorporate all others, RDF allows any number of models to be in use at the same time. Different applications can have different views of the same data, and the addition of a new view does not impact existing views.

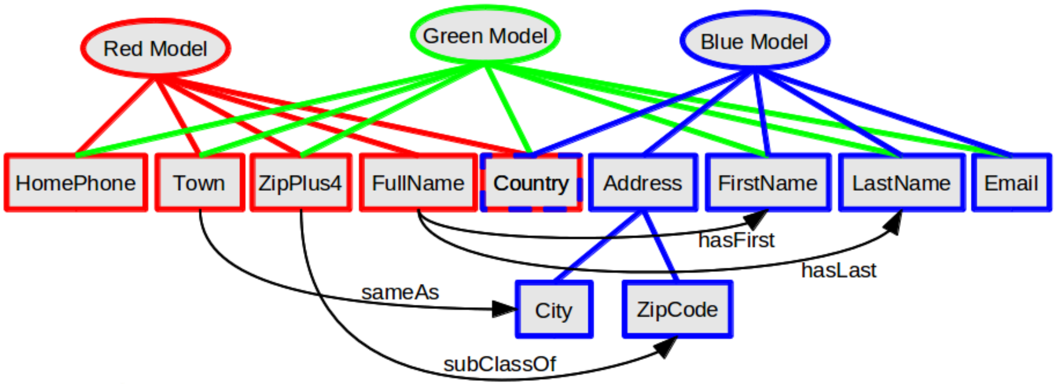
For example, two customer address models – Red and Blue – may have been developed independently for different applications, each one representing similar information in different ways.



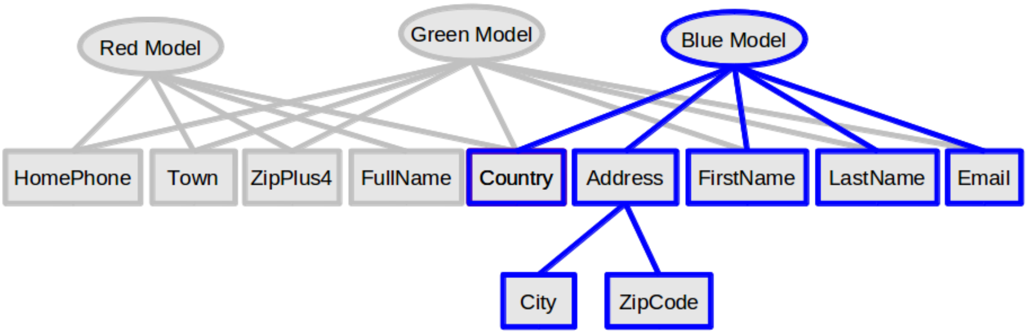
Later, these two models are used together in the same data, and a **bridge model** is used to link them together.



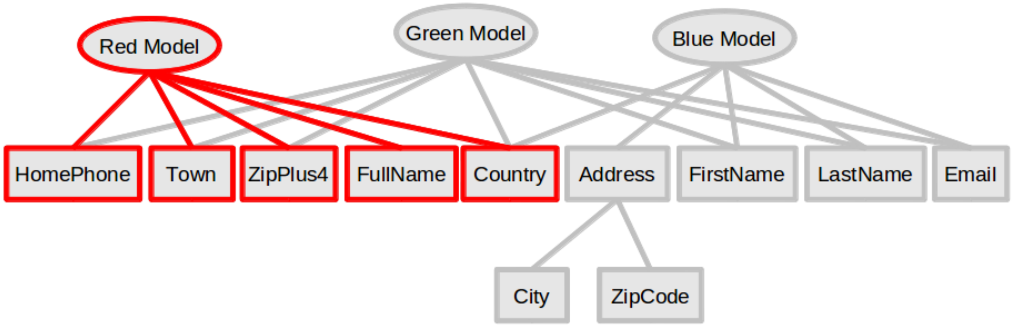
The bridge model adds more RDF relationships (shown in black) to semantically link concepts in the Red and Blue models that are related. Concepts that were already used in common between the two models, such as Country (as determined by use of the same URI), join automatically with no need for a linking relationship. A Green application may later make use of elements from both the Red and Blue models to form its own Green model.



Even though all of these models now coexist, and are semantically linked, each application still retains its own preferred view of the data. The addition of other data models does not break existing applications. The Blue application still sees the Blue model,



the Red application still sees the Red model,



and the Green application sees the Green model, which happens to make use of portions of the Red and Blue models.

Furthermore, by exploiting the semantic links between models, information that originated in one model may be automatically transformed to a second model for use by applications that expect the second model, provided that bridge transformations are available. Some transformations are easy. For example, a 9-digit zip code from the Red model (ZipPlus4) is already a kind of Blue model zip code (ZipCode), since the additional four digits are optional in the Blue model zip code: every ZipPlus4 is a ZipCode (but not vice versa). Other transformations may depend on multiple data elements. For example, to create a 9-digit ZipPlus4 postal code, information from ZipCode, Country, City and Address must be used. Transformations can also draw upon information from multiple models at once. In some cases a transformation may require the addition of external information, if the transformation requires more information than existing models have captured.

This ability to simultaneously accommodate any number of semantically linked models is crucial because it allows new data models and vocabularies to be continually incorporated, without breaking existing software.

Emphasis on semantics

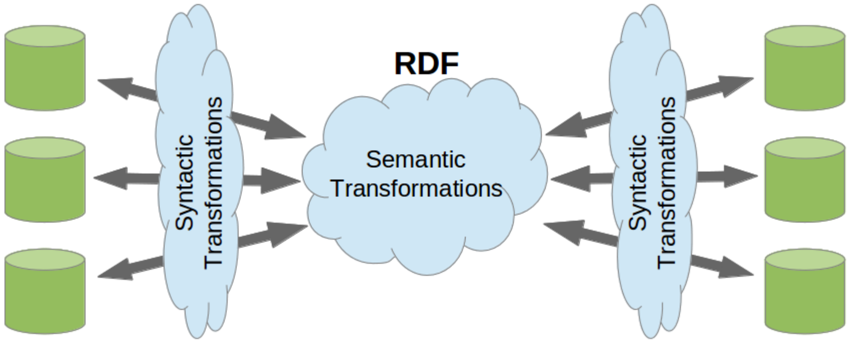
RDF is **syntax independent**: RDF data expresses an abstract model of the information content that is independent of its serialization. There are multiple serialization formats (or syntaxes) for RDF. The same exact information content (or RDF abstract model) can be serialized in different ways. Four of the most popular serializations are N-Triples[9], which serializes RDF as a very simple list of subject-verb-object *triples*; Turtle[10] which is a more compact form that is easier for humans to read and write; JSON-LD[11], which is a more recent JSON-based format; and RDF/XML[12], which is an older XML-based format that is now mostly only used for historical reasons, because other formats are now easier to understand and use.

Being syntax independent, RDF places an emphasis on the meaning of the information rather than on its syntactic representation. This also means that any data format with a suitable mapping to RDF can be used as RDF – regardless of whether it was originally designed to be an RDF serialization. This is significant, because it means that existing document formats can be treated as specialized RDF formats and used along with other RDF content. Existing standard formats do not need to be discarded and reinvented in RDF. Instead, a transformation can be defined from the existing format to the RDF abstract model.

Data transformations

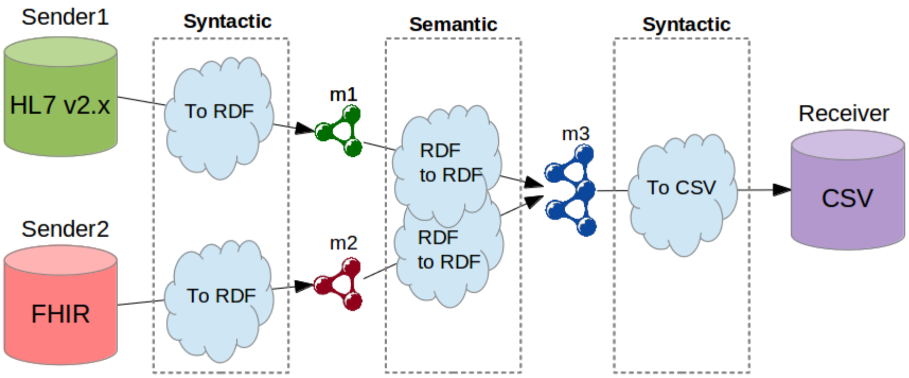
RDF does not by itself give us semantic interoperability. Rather, it acts as a common language substrate for enabling semantic interoperability. If data originates in one form, but is required in a different form, there is no way around the need to transform the data from one form to the other. For the moment we will ignore details of *where* and such transformations will be performed -- they could be done by senders, receivers, intermediaries, or some combination thereof -- and focus only on *what* needs to be done and how it can be achieved.

To simplify the task of creating these data transformations is it helpful to conceptually subdivide them into *syntactic transformations* and *semantic transformations*. The syntactic transformations are used to convert between existing data formats and RDF. The semantic transformations are used to convert between models and vocabularies within RDF. This conceptual division of responsibilities allows the semantic transformations to be potentially reusable across different data formats, and it allows the semantic transformations to be as simple as possible, because they do not need to deal with the idiosyncrasies of the data formats.



Syntactic transformations

To illustrate the syntactic transformations in our example of Sender1, Sender2 and Receiver, HL7 v2.x data from Sender1 must be transformed into RDF. This transformation is not intended to yield the final RDF that will be needed, but transforms the data into an RDF version of Sender1's **native model**, m1, which corresponds directly with the information content of the HL7 v2.x data, without attempting to map to a particular destination model. Model m1 uses the same vocabulary as the original HL7 v2.x data and directly reflects its existing information model (ignoring superficial syntactic details). Similarly, FHIR data from Sender2 is transformed to an RDF version of Sender2's native model, m2.



There are two reasons for transforming first to this native model instead of attempting to transform all the way to the destination model or even to a common intermediate model. The first is that it simplifies the transformation: it can be implemented without knowledge of any other models. This also allows these syntactic transformations to be performed using generic tools. For example, the W3C has defined a generic Direct Mapping from relational databases to RDF.[13] The result of the W3C Direct Mapping is RDF that directly reflects the information model of the relational database to which it is applied.

The second reason for transforming first to the native model is that it avoids tightly coupling the syntactic transformation to any particular destination model, and this allows the transformation to be more reusable. If a new destination model is desired, the same syntactic transformation can be used, and only a new semantic transformation will be needed. It also better insulates the semantic transformations from changes to the syntactic transformation that are the result of changes to the original data format.

As an example, suppose Sender1 sends an HL7 v2.x message looking something like the following (simplified for clarity):

```
OBX|1|CE|3727-0^BPsystolic, sitting||120||mmHg|
```

A syntactic transformation may convert this to RDF expressed in model m1 like the following (written in Turtle, with namespaces omitted for brevity):

```
d1:obs042 a m1:PatientObservation ;
 m1:code ″3727-0″ ;
 m1:description ″BPsystolic, sitting″ ;
 m1:value 120 ;
 m1:units ″mmHg″ .
```

Similarly, Sender2 may send a FHIR message looking something like the following (simplified for clarity):

```
<Observation xmlns="http://hl7.org/fhir">
 <system value="http://loinc.org/"/>
 <code value="8580-6"/>
 <display value="Systolic BP"/>
 <value value="107"/>
 <units value="mm[Hg]"/>
</Observation>
```

A syntactic transformation may convert this to the following RDF:

```
d2:obs-091 a m2:Observation ;
 m2:system "http://loinc.org/" ;
 m2:code "8580-6" ;
 m2:display "Systolic BP" ;
 m2:value 107 ;
 m2:units "mm[Hg]" .
```

Syntactic transformations are also used in the opposite direction: to serialize to a particular required data format after semantic transformations have been performed. In the diagram above, the transformation would be from a third information model m3 (discussed next) to a comma-separated-values (CSV) format required by Receiver.

Semantic transformations

Although the syntactic transformations have converted Sender1 and Sender2's data into RDF, the data is expressed in models m1 and m2 that Receiver will not understand, because Receiver understands only model m3. Therefore, semantic transformations are performed from RDF to RDF to transform from source models and vocabularies to destination models and vocabularies. The purpose is to achieve **semantic alignment**: to express all of the information in the same, desired information model and vocabularies – in this case m3 – so that the information can be meaningfully combined and understood by the Receiver.

There are many ways these semantic transformations can be performed. Here is an example of one technique, which uses a SPARQL[14] query as a transformation rule to convert from Sender1 model m1 to model m3 (namespaces omitted for brevity):

```
# Transform m1 to m3
CONSTRUCT {
 ?observation a m3:Observation ;
  a m3:BP_systolic ;
  m3:value ?value ;
  m3:units m3:mmHg ;
  m3:position m3:sitting . }
WHERE {
 ?observation a m1:PatientObservation ;
```

```
  m1:code ″3727-0″ ;
  m1:value ?value ;
  m1:units ″mmHg″ . }
```

Here is the transformation from model m2 to model m3:

```
# Transform m2 to m3
CONSTRUCT {
 ?observation a m3:Observation ;
  a m3:BP_systolic ;
  m3:value ?value ;
  m3:units m3:mmHg . }
WHERE {
 ?observation a m2:Observation ;
  m2:system ″http://loinc.org/″ ;
  m2:code ″8580-6″ ;
  m2:value ?value ;
  m2:units ″mm[Hg]″ . }
```

How can a system automatically know what semantic transformations to apply? Metadata, also expressed in RDF and carried with the data, can indicate what data models and vocabularies are used. By also knowing what data models and vocabularies the Receiver expects, a system can automatically deduce what semantic transformations are needed. For example it may determine that it needs to transform from m1 to m3, or from m2 to m3. It could then lookup the appropriate transformations in a catalog and apply them. Such transformations might not be performed in a single step, but may involve a short series of transformations forming a **transformation path** from source model(s) to destination model.

Incorporating non-standard concepts

RDF permits both standard and non-standard vocabularies to be used and intermixed, using the techniques described above. This provides the best of both worlds: standard vocabularies can be used to whatever extent they are available, thus simplifying transformation processes and enabling semantic interoperability, while non-standard or emerging vocabularies can be used whenever they are needed, and can be linked with standard vocabularies. This allows the adoption of standard vocabularies to proceed gracefully at whatever pace is possible.

One may wonder why it is important to accommodate vocabularies (or individual concepts) that have not yet been standardized. If a concept has not been standardized, what good will it do to include it in the transmitted data? How will anyone know how to interpret it? The answer is that the adoption of new concepts does not happen all at once by all parties. Some parties will be able to make use of a new concept even before it is standardized. Indeed, they have a business incentive to do so. And if Linked Data principles are used, then each concept's definition can be easily obtained by dereferencing the concept's URI, this making it easy to bootstrap the use of new concepts. The inclusion of non-standard concepts also smooths the path toward their eventual standardization, by allowing interested parties to gain experience with them.

However, the ability to accommodate non-standard concepts is both a blessing and a curse. The downside is that it could allow healthcare providers to use non-standard concepts *even when standard concepts are available*, which would impede semantic interoperability rather than enabling it. And unfortunately, since healthcare providers in the USA have no natural business incentive to make their data understandable by others (especially their competitors), incentives must be provided in other ways to ensure that standards are used whenever possible. Therefore, **carrot and/or stick incentives must be enacted to encourage the use of designated standards**.

The problem of proprietary vocabularies

High-quality healthcare vocabularies cost money to create and maintain. Hence, some have restrictive licensing requirements that prevent them from being freely used. This creates a barrier to their use. To best enable semantic interoperability, this barrier should be reduced or eliminated.

The aspect of this barrier that is most important to reduce or eliminate is the barrier to use (or *read*) and understand data that was expressed using a proprietary vocabulary. While one party may choose to gain the benefits of using a proprietary vocabulary when capturing (or *writing*) its healthcare data, another party that needs to receive and understand data from the first party should not be forced to sign or pay for a license to that vocabulary in order to properly interpret and use the received data.

Policymakers can reduce these barriers by encouraging or requiring the use of vocabularies that are free and open for use in reading – and ideally also writing – healthcare information.

Recipe for semantic interoperability

In the exchange of healthcare information, semantic interoperability cannot be realistically achieved all at once in a "big bang" fashion. But it can be realistically achieved on a progressive basis in which the scope of semantic interoperability is as large as possible and continually increases as more concepts become standardized. This section outlines key principles for achieving this goal.

1. Exchanged healthcare information should be **machine-processable**, as structured data, whenever possible. *This is important to enable useful, automated machine processing.* Narrative or other unstructured information should also be included as an adjunct to convey additional information that cannot be adequately conveyed as structured data.

2. Exchanged healthcare information should be in an **RDF-enabled format**, either: (a) in a standard RDF format directly; or (b) in a format that has a standard mapping to RDF. *This is important to simplify information transformations, by adopting a common language substrate.*

3. Designated **standard vocabularies** should be used whenever possible, i.e., for all

concepts that are defined in designated standard vocabularies. *This is important to simplify semantic interoperability.*

4. The set of designated standard vocabularies should be **continually expanded and updated**. *This is important to continuously increase the scope of semantic interoperability.*

5. ***All* requested information** should be provided in an RDF-enabled format – not only those concepts that are available in designated standard vocabularies. For concepts that are not (yet) defined in designated standard vocabularies, best available vocabularies should be used. *This is important to enable the smooth evolutionary adoption of new concepts and to ensure that all potentially useful information is available to authorized requesters.*

6. Existing standard healthcare vocabularies, data models and exchange languages should be leveraged by defining **standard mappings to RDF**, and any new standards should have RDF representations. *This is important to preserve investments in existing and future standards.*

7. Exchanged healthcare information should be **self-describing**, using Linked Data principles, so that each concept URI is de-referenceable to its definition. *This is important to ensure consistency of interpretation and to efficiently bootstrap the adoption of new concepts.*Healthcare information should be exchanged using **free and open vocabularies**. *This is important to prevent business and legal barriers from impeding the beneficial exchange and use of healthcare information.* Note that this would not prevent healthcare providers from using proprietary vocabularies internally.

8. Adequate **incentives for semantic interoperability** should be enacted, to ensure adherence to the above principles. *This is important to overcome the lack of natural business incentive to achieve semantic interoperability in the healthcare industry.*

Proving feasibility

Many who have worked with RDF and related semantic web technologies believe that this approach to achieving semantic interoperability is viable.[15] All of the features of this approach have been applied individually in healthcare, the life sciences or other areas, but the totality of this approach has not been proven to work on the scale needed for nationwide (or worldwide) adoption. To demonstrate viability on a nationwide scale, research should be undertaken to do the following:

- Build a working, networked system – a reference implementation – that demonstrates this approach on a concrete end-to-end use case involving at least three parties – two senders and one receiver – in which healthcare information about the same patient is received from both senders and usefully combined and used by the receiver.
- Demonstrate the feasibility of all important features of this approach, including:
    - syntactic transformations

    - semantic transformations
    - selecting and applying semantic transformations
    - incorporating a new vocabulary and deprecating an old vocabulary (including how it affects semantic transformations)
    - hosting concept definitions
    - privacy and security adherence
- Design and run stress tests that simulate the adoption of this approach on a nationwide scale, to identify important scaling issues.
- Recommended a set of conventions that would facilitate adoption, such as conventions for identifying semantic transformations, for requesting RDF healthcare information, etc.

Conclusion

The successful adoption of a universal healthcare exchange language is more difficult than it may seem, but it is both feasible and worthwhile. The best available candidate is RDF, primarily because it enables multiple data models to be used together and semantically linked. The viability of this approach for achieving semantic interoperability on a nationwide scale should be proven by building, demonstrating and stress-testing a working reference implementation.

References

1. President's Council of Advisors on Science and Technology: Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward. Executive Office of the President, December 2010. http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf (Accessed 25-Aug-2013)
2. Heather Ryan, Angela Trosclair and Joe Gfroerer: Adult Current Smoking: Differences in Definitions and Prevalence Estimates – NHIS and NSDUH, 2008. Journal of Environmental and Public Health, Volume 2012 (2012), Article ID 918368, 2012. http://dx.doi.org/10.1155/2012/918368 or http://www.hindawi.com/journals/jeph/2012/918368/ (Accessed 25-Aug-2013)
3. W3C: Linked Data. http://www.w3.org/standards/semanticweb/data (Accessed 25-Aug-2013)
4. T. Berners-Lee, R. Fielding, L. Masinter: Uniform Resource Identifier (URI): Generic Syntax. IETF Request For Comments (RFC) 3986, January 2005. http://www.w3.org/standards/semanticweb/data (Accessed 25-Aug-2013)
5. Health Level Seven International: HL7 Standards – Master Grid. http://www.hl7.org/implement/standards/product_matrix.cfm?ref=nav (Accessed 25-Aug-2013)
6. Health Level Seven International: Fast Health Interoperability Resources (FHIR) v0.11. http://www.hl7.org/implement/standards/fhir/fhir-book.htm (Accessed 25-

Aug-2013)

7. Wikipedia: Comma-separated values. http://en.wikipedia.org/wiki/Comma-separated_values (Accessed 25-Aug-2013)
8. Graham Klyne, Jeremy J. Carroll, Brian McBride: Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-concepts/ (Accessed 25-Aug-2013)
9. Gavin Carothers and David Beckett: N-Triples, A line-based syntax for an RDF graph. W3C Working Group Note 09 April 2013. http://www.w3.org/TR/n-triples/ (Accessed 25-Aug-2013)
10. Eric Prud'hommeaux, Gavin Carothers, David Beckett, Tim Berners-Lee: Turtle, Terse RDF Triple Language. W3C Candidate Recommendation 19 February 2013. http://www.w3.org/TR/turtle/ (Accessed 25-Aug-2013)
11. Manu Sporny, Gregg Kellogg and Markus Lanthaler: JSON-LD 1.0, A JSON-based Serialization for Linked Data. W3C Last Call Working Draft 11 April 2013. http://www.w3.org/TR/json-ld/ (Accessed 25-Aug-2013)
12. Dave Beckett and Brian McBride: RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-syntax-grammar/ (Accessed 25-Aug-2013)
13. Marcelo Arenas, Alexandre Bertails, Eric Prud'hommeaux, Juan Sequeda: A Direct Mapping of Relational Data to RDF. W3C Recommendation 27 September 2012. http://www.w3.org/TR/rdb-direct-mapping/ (Accessed 25-Aug-2013)
14. Steve Harris, Andy Seaborne, Eric Prud'hommeaux: SPARQL 1.1 Query Language. W3C Recommendation 21 march 2013. http://www.w3.org/TR/sparql11-query/ (Accessed 25-Aug-2013)
15. David Booth, Charlie Mead, Michel Dumontier, Tracy Allison Altman, Rafael Richards, et al: Yosemite Manifesto on RDF as a Universal Healthcare Exchange Language. Position statement from the 2013 Workshop on RDF as a Universal Healthcare Exchange Language, San Francisco. http://yosemitemanifesto.org/ (Accessed 25-Aug-2013)