The Linked Patient

Patient-Centric, Health-Data Publication

The 2010 PCAST Report on realizing the full potential of health information technology ...

> "**current approaches** to [Patient] data exchange and aggregation, which are often bilateral or document-based, **do not**, in our view, present a clear path to scalable national solutions that would **trigger transformative innovation** and use of health IT. In this sense, there is potentially a large gap between the current path and the potential for IT to improve health and healthcare"
>
> -- PCAST

so that

> "the Federal Government (should) **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"
>
> -- PCAST

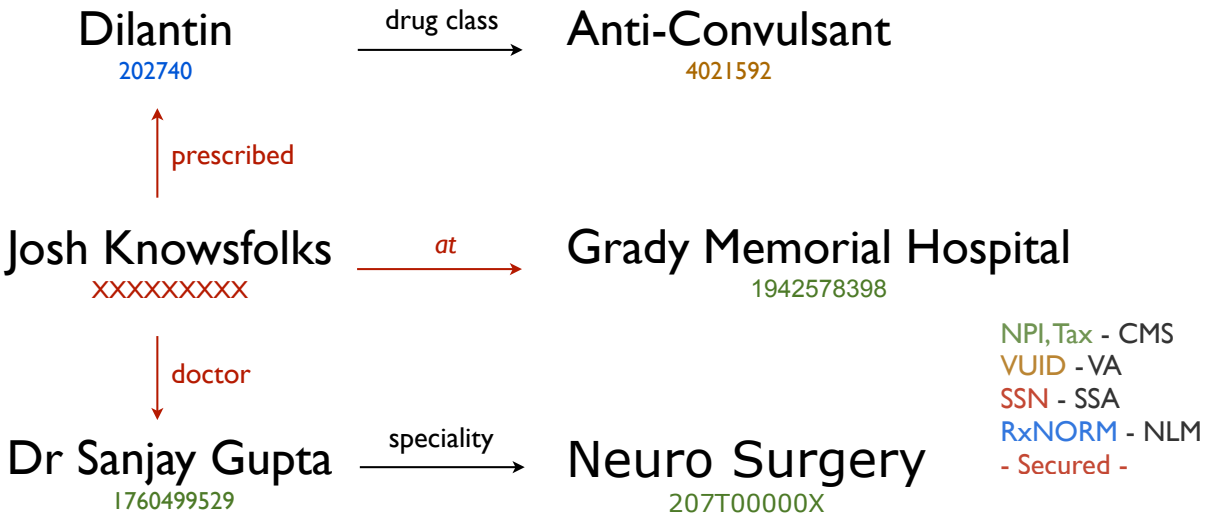
This document will show that **Linked Data**, a standard, web-centric approach to data publication answers PCAST's call.

Linked Data fits Health-care

"Information is the lifeblood of modern medicine, [and] health information technology is destined to be its circulatory system" -- David Blumenthal

Any description of **patient** care - diagnoses, prescriptions, procedures - draws in **institutions** like hospitals, pharmacies and doctors and medical **know-how** like the class and makeup of a drug and its interactions with others.

Josh Knowsfolks was prescribed the anti-convulsant, Dilantin, at Grady Memorial Hospital by Dr Sanjay Gupta, a Neurosurgeon.



Notice that a patient's care links many things and that thanks to billing and licensing requirements those things are unambiguously identified with government issued identifiers - HHS gave Sanjay Gupta the national provider identifier (NPI), 1760499529, a handle that will follow him for life. How long will Twitter's "sanjayguptaCNN" last?

And linking and identifying goes on from here. Where is Grady Memorial Hospital? In the zip code, 30303, according to the postal service. Dilantin is contra-indicated for the drug Warfarin which the VA gives the identifier, 4018548, has the active ingredient, Phenytoin, which the NLM assigns RxNorm 8183 and treats Epilepsy, which gets MeSH identifier M0007580, again from the NLM.

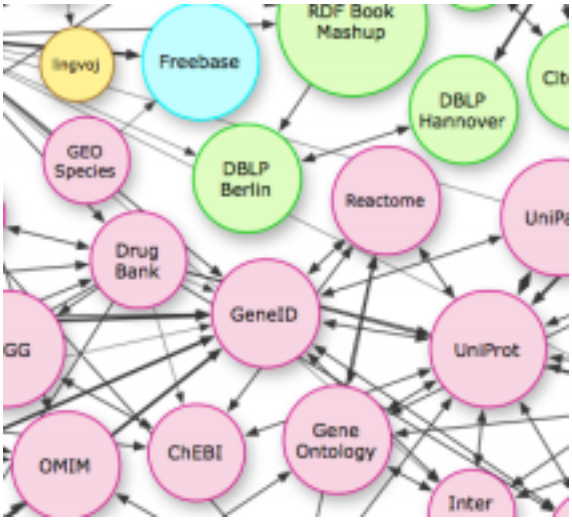
From a data perspective, Health-care requires description of a huge volume of highly interlinked, clearly identified things, some private, most open, making it an ideal domain for the W3C's **Linked Data** approach to data management. Health-care is a world of linked things - the data that describes it should be linked too.

Beyond Google Guesses

Point your browser to the URI *http://cms.hhs.gov* and up comes a web page, a document designed for you, the human. However, this arrangement of data - images and blocks of text designed to be easy on the eye, links in context that point to other information - is difficult for machines to understand. In a word, machines are dumb - they require a lot of structure in their data.

One route to machine understanding is clever algorithms that pick apart web pages designed for people. This is the approach taken by Google and other search engines but it is a hit and miss affair. Another is to provide a machine-friendly version of the information on a web page. By adding support for data for machines alongside a page for humans, an approach called “Linked Data”, you remove the need for cleverness and work-arounds.

Defined by the W3C, the international standards body for the web, Linked Data builds on the web’s foundations - URIs, those identifiers that start with “http://” and uniquely locate things, HTML, a format for human-readable, linked-information, HTTP, the web transmission protocol, web security, used everyday by bank and other sensitive sites - and adds RDF, a machine-friendly data format, as well as extensions to HTML so that machines are as welcome on the open web as people.



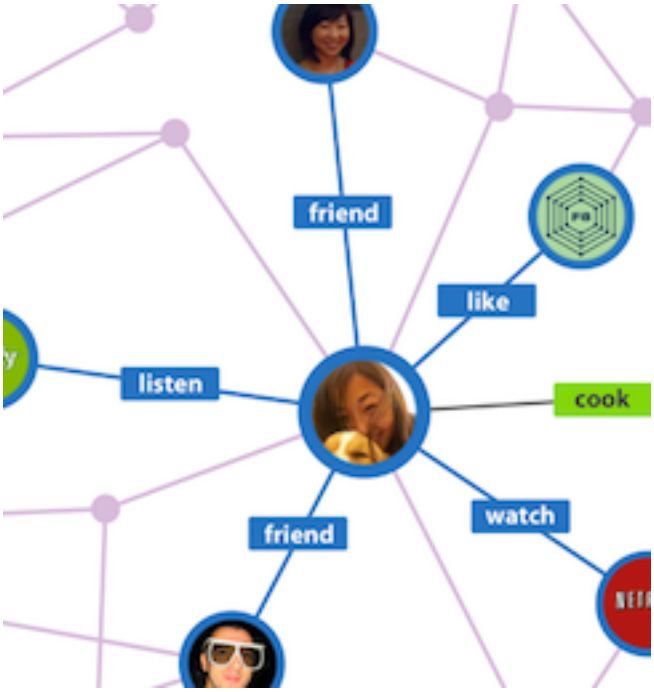
With these modest additions to our most scalable computer architecture, man and machine can browse in harmony, fetching what they need, traversing links. Where man sees a page in his browser, machine gets to process easy-to-understand data.

Just as it hosts trillions of human-readable pages, Linked Data allows the web to host data on any scale, each piece uniquely identified. Health-care presents huge numbers of linked things, many of which have unique identifiers already. It’s as if Linked Data was created to describe it but progress has been slow. Though Linked Data is used to publish medical know-how, **patient information remains off this securable web**.

Facebook took advantage

Facebook says ...

> at Facebook's core is ... people and the connections they have to everything they care about.



In the real world, people link to things - other people, music, news reports, recipes ... - in particular ways - friend, listen, watch, cook ... and Facebook's data management reflects this.

Facebook's platform is web-based but the traditional "web of documents" wasn't sufficient. They needed the web to go beyond linking human-readable documents to host data about highly linked things in an easy to process form. Sound familiar?

For example, the review site Yelp publishes a page about Dr Sanjay Gupta with his address, a photograph and a series of reviews. Facebook doesn't want to guess which information represents an address, what's a review, which image is a doctor's photograph rather than Yelp's logo. It wants Yelp to serve up easy to process structured data with the same information as its human-readable web page.

Facebook didn't need to develop this web overlay themselves. Internally, they implemented pieces of W3C's Linked Data approach in what they called the Open Graph Protocol and documented how traditional web page publishers could publish Facebook-friendly Linked Data.

Given Facebook's influence, Linked Data is finding its way onto the traditional web with publishers from the New York Times to small recipe sites adding machine-friendly, structured data at every URI and a growing ecosystem of data-centric applications, both academic and commercial has sprung up.

Grady Memorial, the patient-data master

Let's visit a brave new world. As now, Medical Providers are told to share patient data with each other and with their patients including Mr Knowsfolks. Grady Memorial has all the data they need but it's in a custom, internal EHR system. How should they publish it? In this new world, HHS, the "Facebook of Health-Data", shows how with its "Guide to Linked-Patient-Data" and a full demonstration hosted on HealthData.gov.

Grady must support two types of user - the likes of Josh who wants to browse his medical information in the same way he banks online ("Patient Portal") and the medical application running in another one of his doctor's offices that wants this or that piece of information ("automated Health-Information Exchange"). The guide tells Grady how to follow W3C guidelines to provide pages for Josh and data for medical applications.

Because this is the web, the hospital doesn't need to host every piece of data itself. Other publicly available sites publish the open parts of Josh's medical data.

On *HealthData.gov*, HHS provides Linked Data access to its national provider index. The description of Grady itself and Dr Gupta are hosted there at two NPI-identified locations - *http://health.data.gov/id/npi/1942578398* and */1760499529*. HHS also exposes its provider taxonomy as Linked Data with "Neurosurgeon" at *http://health.data.gov/id/tax/207T00000X*.

Grady leverages drug data hosted publicly by Caregraf. Locating Dilantin's information is easy. NIH gives it the identifier, rxnorm:202740, so Caregraf hosts its data at *http://datasets.caregraf.org/rxnorm/202740*.

But what about the private information? Josh's prescription is sensitive. He wouldn't want anyone but certain doctors to know that he is taking something for Epilepsy, no more than he'd want people to know his bank balance. While identifiers and details for the context of the prescription - which drug, which doctor, which hospital - can be open, access to the prescription itself must be strictly controlled. In this Grady is as security conscious as his bank.

This is web data so the prescription itself gets a URI too, *http**s**://www.gradyhealth.org/patients/prescriptions/110133*. Notice that "**s**" - this is secure HTTP, the sort used by Josh's bank. Only permitted parties, logged in, can see the prescription. What's more, Grady's web server implements access-control on its data. Not every party logged in can see Josh's information, only permitted doctors and their systems and of course Josh himself.

When Josh logs into Grady's Patient Portal, he can call up his prescription in his browser. If he wants to know more about Dilantin, he clicks on a link and is taken to the Caregraf site and on he goes from there.

The system of Josh's other doctor logs in and requests his current prescriptions. As it is a system, it gets a machine-friendly form of the same information seen by Josh. Earlier this system had pulled more of Josh's prescriptions from another doctor's office. Because all the data retrieved is structured, it is easy to merge and that's just the beginning.

This system always checks if drugs may interfere with one another, easy because it can follow links that describe Dilantin and other drugs Josh is taking and at those links it will find structured information about drug interactions. It notices Josh is suffering from sleeplessness. Is this a side effect of one of these medications? Perhaps FDA drug data can help and off it goes to a site which has the latest information on each of Josh's medications in structured form.

Thanks to clear HHS guidance, both Josh and machine get information in the form they need, security is preserved, Grady doesn't do all the work - links to other sites delegate much of the information load - and the technologies used are easy to understand. Grady's I.T. staff find nothing strange about the web and some have developed novel applications around this easy to understand data.

Of course, what's above is a fantasy world. Grady would never publish patient data this way. Why? Because HHS not only doesn't suggest Linked Data for patient data exchange, it effectively precludes it. It says patient data is special data, so different that it needs a custom infrastructure. In short, **HHS says no** to patient data on the web, that **Health-care will go it alone**.

HealthData.gov - the Health-Data conductor

"When information is brushed up against information,
the results are startling and effective."
-- Marshall McLuhan

"HealthData.gov is a public resource designed to bring liberated health datasets, innovation challenges, and applications and tools to the public to help increase public knowledge and solve problems in health."
-- Todd Park, United States Chief Technology Officer

We are fortunate that much of the Health-care **know-how** and **institutional information** in the United States is under one umbrella, HHS, and that HHS is openly publishing this information as machine-processable datasets on HealthData.gov.

Currently, this site presents a catalog of datasets, organized by agency. Most publication is "**open download**" - datasets come in files with custom formats and can be freely downloaded. Then it is up to every user to break these files down and make their contents usable.

Unfortunately given so much information, there is no indication of how these datasets fit together - the linked nature of health-care is not reflected - and there is no indication of which sets represent axes around which others gather.

The next challenge for HHS is to move from hosting **stand-alone datasets for download** to orchestrating their **publication as Linked Data.**

As shown above, a description of any clinical event directly involves know-how and institutions described in HHS datasets. The more HHS datasets are coherently published as Linked Data, the easier it will be for hospitals and others to publish patient information on the web and focus scarce resources on novel applications.

The viability of linked-patient-data relies on the success and further development of this site both as a destination in itself and as a point of orchestration for others that publish HHS data.

Not all identifiers are created equal

A National Provider Identifier (NPI) is a unique 10-digit identification number issued to health care providers in the United States by CMS. All individual HIPAA covered healthcare providers or organizations must obtain one. The NPI is the way to identify the "institutional" things involved in a patient's care - Sanjay Gupta has one and so does any hospital he practices in.

It is an obvious focal point for HHS data about an institution, an ideal way to organize information from CMS Hospital Quality Data to Medicare cost data to FDA reports on Mammography facilities. Were there a picture of the links between HHS datasets, many would point into the NPI.

Today you can download NPIs as a blob, CMS provides access through a semi-secured web interface to NPPES, its NPI management system, and some outside medical billing sites provide custom access but none provide Linked Data and none draw in other datasets.

In this, the NPI dataset is like all but one other on HealthData.gov - the exception is CMS Quality Measures whose Linked Data publication (http://health.data.gov/def/hospital/Hospital) points the way for the NPI dataset - both people (HTML) and machines (RDF) can browse Quality Measures in detail. As a point of integration, NPI data deserves the same treatment.

Orchestrate creative publication

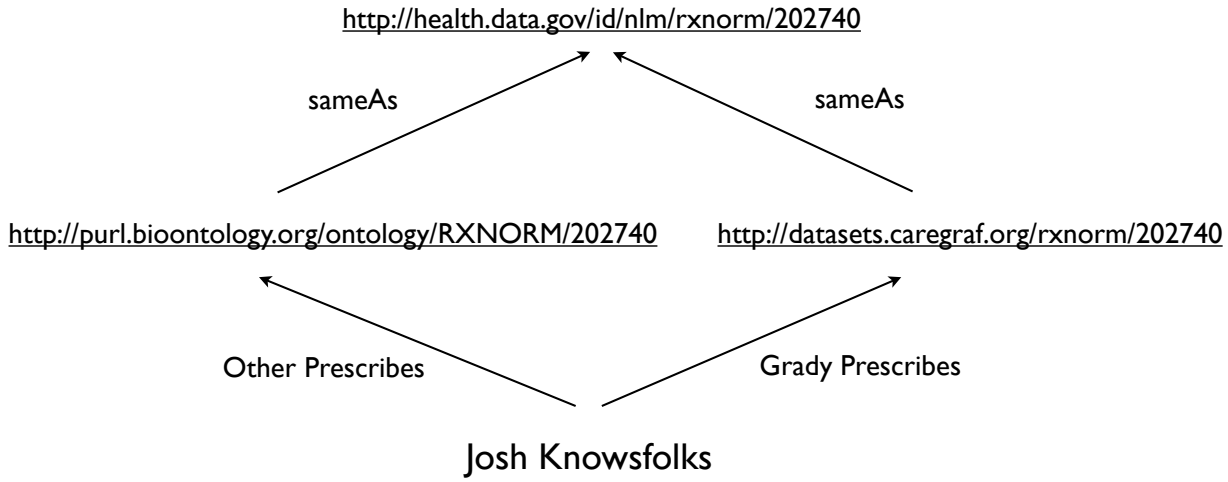
Having provided clear access to its datasets, must HHS itself now instigate and finance granular publication of them all? In many cases, no and some of these datasets have been published elsewhere already.

For example, Stanford's national center for bio-medical ontology publishes many HHS clinical and bio-chemical datasets on a Linked Data site (bioportal.bioontology.org/). This work is being funded by HHS - "the Center is funded by the National Institutes of Health (NIH) and is part of the network of National Centers for Biomedical Computing". Our company, Caregraf, hosts VA, FDA, NIH and other clinical datasets on an open Linked Data endpoint (datasets.caregraf.org).

Overlap between Caregraf and Bioportal highlights a key attribute of data on the web - the same information may be represented in different ways by different parties and be linked out to different information too. Different uses favor different arrangements for data and best practices are still evolving.

Of course the danger with multiple publishers and forms is chaos. Delegated publication needs some organization and organization on the web revolves around URIs.

Consider two Dilantin prescriptions for Josh, one by Grady and a second from another facility.



Here Grady uses Caregraf's Linked Data service to identify its drugs but the other facility uses Bioportal's equivalent. Pity the poor dumb application trying to digest these prescriptions - one type of drug, same NLM identifier (202740) but two URIs.

Enter **sameAs**. With this simple Linked Data technique, both Caregraf and Bioportal can tell an application that their URIs identify the same thing just by sameAs-linking to an **authoritative equivalent hosted by HHS**. Through these links, any application would know it faces two prescriptions for the same drug.

Such Authoritative HHS URIs need to be set in stone and follow a consistent scheme - agency mneumonic/identifier? But though they need to resolve, they don't need to return more than the name of an identified thing.

For example, http://health.data.gov/id/nlm/202740 could just return the label "Dilantin" and various details about Dilantin could be hosted at many different locations by many parties. Their primary purpose is to **coordinate publication of data about the same things by multiple outside parties**.

Must HealthData.gov host these URIs? Yes. Facebook is a center of gravity of much of the web because it owns the URIs of so many people. Google Maps owns physical space for the web for the same reason. Authoritative URIs for HHS identifiers, in particular for focal points like NPIs, will anchor web-resident health data and belong on HealthData.gov.

The private sector should innovate around data representation, linking and analysis but these government-sanctioned points of coordination should not belong to a private data broker. To put it figuratively - let everyone play but provide a conductor.

The Wrong Messages from HHS on Patient Data

*"in operational and practical fact, **the medium is the message**. ... the personal and social consequences of any medium ... result from the new scale that is introduced into our affairs"*
-- Marshall McLuhan, Understanding Media

Today **patient data** is treated very differently from its peers, health-care **know-how** and **information on institutions**. For a variety of reasons, HHS chose to promote a unique medium for this data which makes its publication, analysis and presentation overly complex and has greatly lessened any chance of patient-data exchange transforming health-care.

While some special handling is merited - access to patient data must be closely controlled - there is no technical difference between patient data and its peers and all three types of health information - know-how, institution and patient - **should populate the same data space**.

"Send over his file"

Automation usually starts in the real world - computer screens become "desktops", you post things on your Facebook "wall", you "browse" the "web" - and a well chosen metaphor advances automation. Choose badly however and much of your automation will be academic.

Doctors talk about patient "records" or "files" which made the "patient's file" the obvious motif for patient-data exchange. However, unlike the "web" or the "wall", this "file" has many technical drawbacks. It is a self-contained blob of everything known, transferred in bulk. Blobs get digested whole unlike granular information which is chosen and processed in pieces.

As the PCAST report said ...

> "current approaches to data exchange and aggregation, which are often bilateral or document-based, do not, in our view, present a clear path to scalable national solutions that would trigger transformative innovation and use of health IT. In this sense, there is potentially a large gap between the current path and the potential for IT to improve health and healthcare"

and suggests ...

> "the best way to give clinicians a unified, patient-centric record tailored for each medical encounter is to store, maintain, update, and exchange the data as small, distributed, metadata-tagged elements"

In other words, **the file is the wrong motif for patient data exchange**. Providers should publish granular, networked data and let applications pick and chose what they need. The web? Linked Data?

For PCAST, a roll-up of everything known should not be the mechanism for exchange. Instead, reports of all sorts are products of data-assembling applications ...

> varied assembly into documents or reports can itself be a robust, entrepreneurial marketplace of applications.

Every Medium sends Messages. The self-contained patient file cries out that "patient data is intrinsically special and so needs its own standards", "no off-the-shelf, general purpose web technology could suffice for something this special", "don't worry about the likes of HealthData.gov - that's for a different sort of data, nothing to do with the patient directly".

Of course no one explicitly wrote or said such things. Patient data sequestration wasn't anyone's intent but it was the **unintended consequence**. The medium chosen by HHS gave rise to a series of custom standards for patient data markup, identification and transfer and the rapidly evolving **web now goes on without the patient**.

Unlinked by nature

The web-friendly URI is a truly liberating identification mechanism. Through it, trillions of pieces of information have been published and inter-linked. But the patient files promoted by HHS don't identify with URIs. They use another, older scheme called Object Identifiers (OID).

As web URIs begin with an organization's hostname, an OID begins with their root id. With these naming-bases in hand, an organization can identify anything it wants though here any similarity ends. OID roots are jealously guarded - unlike freely available hostnames, only the special can get one, hardly a formula for devolved data publication.

And of course data that doesn't leverage URIs is intrinsically not web-data. It stands alone, off-line. Put an OID into your browser and you'll be taken to an error page. This is wholly anti-social technology.

Limited Expression

Some patient data is understood by all. In many cases, its details are defined in law - Prescriptions, Radiology tests, Vital sign measurements, ... these are described in much the same way in every institution. As a result it was straightforward to specify one standard vocabulary for such data and specify a file format around it.

But one vocabulary only takes you so far. What about more specialized information captured by a hospital's EHR? For example, the VA's VistA notes if a veteran is an "agent orange victim". That's something a Veteran's outside doctor should know but as this isn't interesting or applicable to everyone, the standard vocabulary can't express it.

PCAST pointed out this problem ...

> "we believe that any attempt to create a national health IT ecosystem based on standardized record formats is doomed to failure"

Linked Data avoids being boxed in - you can publish any information and it is up to clients to ignore what they don't understand. When you publish data, you also publish

the vocabularies it uses. Clients know what they are getting and publishers can define new and ever richer vocabularies which clients can adopt over time.

PCAST again ...

> "ONC might, by standardizing a universal exchange language whose semantics is intrinsically extensible, unburden itself of a potentially never-ending and intrusive government role in the harmonization of health record meanings across all private sector products. An open, extensible language will allow products to compete, balanced with other competitive features, on the basis of the breadth of their abilities to understand multiple semantic realms."

This is "data modeling, web-style", a devolved approach to expression definition that recognizes that Medicine is a world of specialities and each must evolve its own lingua-franca. Again, as-is Linked Data answers PCAST's call.

Cloister to cloister

PCAST made this observation about security ...

> "complex mandates of both HIPAA and state laws and regulations leads organizations to equate protection to sequestration"

Patient Data exchange was meant to address this misconception head on. Data would come out of silo's but security would still be preserved. Data would no longer be sequestered but would still be secure. Unfortunately, **sequestration is still going strong** - a blob moved from one special place to another is still a cloistered data dump.

PCAST wants granular security for granular data ...

> a more sophisticated privacy model, one where privacy rules, policies and applicable patient preferences are innately bound to each separate tagged data element and are enforced both by technology and by law

Linked Data was synonymous with open data - data published with no concern for authentication and access control. Sensitive data publication introduces new demands and the W3C and others recognize this. Though active work is being done, it is one area where HHS may have to step in to select acceptable approaches.

Weening off the File-motif

Despite its drawbacks, despite alternatives, will file-centric, patient data exchange continue for the foreseeable future? When will linked-patient-data take its place?

What's past is prologue. Here's Wes Rishel of Gartner on the demise of OSI Networking ...

> Perhaps the most famous failure of a "second network" was the ISO Open System Interconnect (OSI) protocol suite, which was the planned TCP/IP killer of the early 1990s. The complete replacement for all levels of TCP/IP fixed many known problems such as the paucity of IP addresses.
>
> Dozens of consultants with solid TCP/IP credentials spent years developing it. The DoD decreed that it would no longer buy systems based on TCP/IP, NATO agreed

on it, GM build a whole plant-floor architecture based on it and — guess what? It never happened. Instead we adopted network address translation until IP6 could be rolled out and many other less elegant fixes that could be introduced incrementally.

An advancing “good enough” technology overwhelmed a key, federally mandated and financed mechanism. Sound familiar?

General purpose, web-based data management is advancing rapidly. Patient data is data, highly Linked Data. We feel that the progress of Linked Data as a publication mechanism will overwhelm the federally promoted File-motif. It’s just a question of when.

While HHS could let linked-patient-data evolve at its own pace and continue providing unqualified support for the File-motif, we believe it is better to promote this web-based alternative sooner rather than later. A subtle course would be to publicly explore granular patient data publication while continuing support for the older motif.

Once linked-patient-data has proven itself, it could be given equal status and the market can decide from there - given the option, what would Grady Memorial do?

Health-care **needs disruptive Health-IT** but what has the File motif brought to life? Fortunately, as PCAST notes ...

“the definition of meaningful use under HITECH ... leaves CMS broad discretion”.

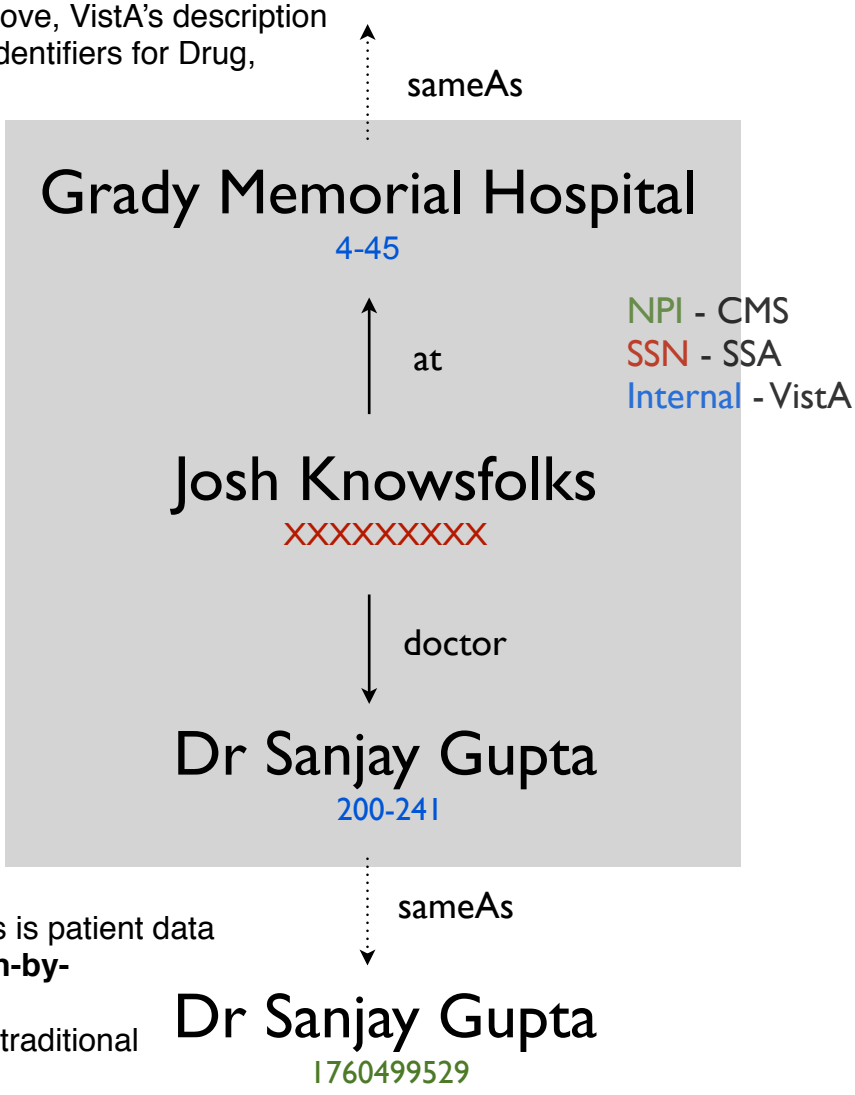
EHRs are Standing By

Providers today are encouraged to capture patient-care in an Electronic Health **Record** system (EHR), a name that suggests a keeper of monolithic per-patient data, data alien to the world of Linked Data. If the systems that hold patient data resist Linked Data, how can hospitals publish it? Fortunately, a look inside mainstream EHRs shows Linked-Patient-Data rather than a gaggle of self-contained records.

Turn the EHR inside out

VistA, the EHR used by the Veteran's Administration is the most widely deployed EHR in the United States. It is mainly used within the VA but it has been deployed in private hospitals too. For our purposes, Grady Memorial is one.

Unlike our descriptions above, VistA's description of a prescription uses its identifiers for Drug, Doctor and Hospital and these identifiers are not understood beyond it. In effect, VistA is its own world, a self-contained linked dataset with no links to the outside.



But publishing linked-patient-data is straightforward - it means linking out those things in VistA with equivalents beyond VistA, isolating data like Drug definitions and Doctor descriptions that are no more than local copies of openly published information and asserting that they are the sameAs publicly available equivalents. This is patient data exchange through **publish-by-linking** and it has many advantages over its more traditional peer, **extract-and-merge**.

One is noted by PCAST ...

> the records themselves can remain in their original locations

By definition, publication doesn't require extraction.

Another clear benefit of publish-by-linking is that it is quantifiable. How many "dead ends" are left? How much progress has been made linking data to the outside? Progress in linking out is reflected in the state of the data at any point in time. Linking-out is a straightforward, easily quantified task - a project manager can't ask for better than that. For a government certifier of Meaningful-Use, **"meaningful" amounts to "how linked"**?

It's been done and it's freely available

Caregraf was involved in a successful Social Security Administration (SSA) project to publish all of the patient data of a real-world VistA as Linked Data.

In this project, the SSA wanted full records of disability claimants from a private sector VistA with two hundred thousand patients. To the SSA, Linked Data publication was incidental - they wanted data rolled up into records - but the process, which is still running, proved that linked patient data publication was both possible and highly efficient. Subsequent projects under TATRC showed the same approach working for RPMS, the IHS's VistA-derived EHR, and on a test system of the VA's.

The publication mechanisms and other Linked Data utilities for VistA are available as open source at http://www.caregraf.org/semanticvista.

HHS, more fuel please

"In short, there is no fuel for an ecosystem of economically self-sustaining healthcare innovation"
-- PCAST

HealthData.gov supplies fuel. Even more would come once HHS ...

- Prioritizes Linked Data publication of focal points like NPI
- Coordinates outside publication by hosting Authoritative URIs for all HHS identified things
- Releases a guide on how to leverage HealthData.gov as the anchor of Health-data publication

The patient needs to join in. HHS should ...

- Release a guide on how to publish patient-care information as Linked Data. Access control for this granular data must be addressed in detail
- Host a linked patient data demonstration showing publication from EHR to both browser and application. We recommend VistA as the EHR because it is an openly available industry bellwether and its patient data has been successfully published as Linked Data
- Repurpose meaningful use tests designed for patient files for use on linked patient data.

About Caregraf

Caregraf, a clinical-data analytics company, helps health-care providers gather and analyze the information they create during the course of a patient's care. The raw material for improved and more cost-effective treatment, such patient-care statements capture not only a patient's condition but also the full workflow of an organization including the extent of its data gathering.

During our many projects, we have created and re-used enabling services, utilities and analyses. We have made the best of these available as open-source because we think Health-IT needs to leap beyond plumbing and data-formatting to focus on in-depth data-analysis, data-centric integration and better representation.

2500 Broadway,
Building F, Suite F-125
Santa Monica, CA 90404
http://www.caregraf.com