NOTES STARTED

OK to Continue Adding Notes until April 26

[Please include Note-taker(s) Names]

Then scroll to bottom & Add your Notes there

****

Day 3 / Session 21 / Space L

Session Title: International Semantic Infrastructure: Requirements for a distributed data economy

Convener: Paul Knowles

Notes-taker(s): Charles E. Lehner

Tags for the session - technology discussed/ideas considered:

Data Management
Data Analytics
Linked Data

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Solutions fit for pandemic? No
MSFT: no problem collecting data from multiple sources, but unable to make sense of the data (semantics)

Inputs and Semantics WG and ToIP
Convened to bring semantics to the Identity Community
OCI
Under this WG, task forces: Storage importability, etc.
Vertical: healthcare task force : underneath it, FHIR focus group

Storage. Rebuild semantics so the data is ready to go.

Linked Data

MIT: David Spivak, Ryan (?)

Collection, storage and exchange.

OCA: Rules. Masking Overlay. Flagged. Some sort of algorithm process of what to do with that flagged data. Rules, algorithms: don't think should be in the storage, but closer to the exchange. Then can simplify OCA - Look at that at exchange side. Character encodings, etc., import. Don't want semantics mixed up, as it complicates the exchange side.
Core overlay need to keep at storage part. Other overlays, more rules-based. Conditional. Rules better off at the exchange side.

Exchange side -> Query languages. Searching data - coming from someone looking for data. Insights-based service provider putting searhc into dynamic data sharing hub, does magic to find data. On exchange side.

Linked Data.

Brent Shambaugh. Integrate computing, processing, storage. Virtual machine to integrate databases. Query language, like Linked Data, over a lot of different systems. SeeQL(?) - translate. Categorical databases (Ryan and David's work): to not really rely on a single source ofo truth, but more rely on transformations between things. Use category theory to have an exact/provable way of doing that. Cat theory has to follow certain rules. Cat theory kind of abstract but provides a framework for unrelated disparate things. Ryan could say how to algebraicly describe things, which would branch off into… Josh at Uber: come up with schemas, get down to the data/logical layer. Different places to go. Way to translate in from out. Might have multiple different ones, want to map from each one, but if you have a vague interpediary in centralized model, loosely defined that both map to, then have a mapping between the two things. Linked Data problem

Paul summarizeing. Work at MIT a combo of linked taa and the query language. Veriy sophisticated. Want to pursue integrating with group. But they are missing this restructure.

Labels
Stack structure
Predefined entries
Link data together -> rich semantics lost. Want to say to JSON-LD people, put back together -> need to rebuild semantics on the data, on the human-readable labels If cross border control want credi to be able to resolve into multiple langauges

Burak says, not necessarily true. In layered architecture. If layers considered to add metadata to linkd data nodes, then don't necessarily lose semantics. Ingest data using layered architecture: can incorporate…

Paul says: because pulling data from multipl einputs, the semantics misalign when you pull them together. That's why need to rebuild the semantics. If you miss the semantic rebuild (I know this from clinical trials, they have this for 20 years same problem). Have to rebuild semantics or will be messy at analytics level. Stopped doing analytics because of the data. Went to OCA. Then back to stats.

Neil: God help you try to edit the bits of the code. Need to know much richer stuff to run the code. Get code, first thing to do? Have to build a mental model of what it looks like. If you don't know the semantics, where it came from, it may be absolute nonsense. If want to put Chinese layer on border agents phone, don't need whole layer. Just the executable version. If want to understand labels...
Linking data together is about machine readability. Involved humans… need to understand. Do it through language. Humans like OCA because can understand data in different languages, makes sense for people. Human element. In that capture space. Want to refine OCA, take out some of the rules parts, masking overlay, conditional overlays, and get it away from OCA as architecture - it convolutes things. OCA only meant for making theings human-readable.

Neil: challenge. Presentation … annotates … data exchange, may have to have rules, data type X in one DB, type X-1 over here. If see this value, talking to database cannot support null. What to do with null? Need to specific place for where these rules are placed in the stack. Can still use layer mechanism to identify what the rules are talking to. Alice going to 5 countries, how to find out what VCs need, what questions will be asked? How to do these queries? Suspect things not searchable through SparQL.

Here's how you can walk through the SparQL thing. Need to know what labs ….

Paul: sounds like a governance thing. If built a specification like shown, would expect Chinese government to specify schema database, and their overlays. In Austrait, they specify their overlays on their side.

Data governance authority. Needs to come from proper authorized place. If come to our country, these are the attributes you need.

Neil: giving enough info in the stack so someone searching from country X can say have it or no. International vocabulary. Somewhere so can retrofit on schema. Query vocabulary layer. Can evolve independently of the data.

Paul: data scientist side different. Can use overlay as Burak might want, to do smarter things. Put rules in overlays, use different schema, maintaining context and predefined entries from the issued source. Would like to go there in the Semantics Group. Stripping out things from OCA that don't need to be in Core. Leave stuff for flexibility of data scientist. Collection side: link data in the wild…
High level approach - build something really special.

Neil: Burak and teams looking through, munging data. If model for how code executable chunk… code executable manipulation. Using the metadata in the stack. Operations not just adding attributes: transformations. How to add active actor to do work other than just translation of JSON-LD.

Burak: OCA has special type of overlays, masking or subset, where they should actually not be overlays, should be operations. Layered schema approach is different: treat schema layers as pure metadata coming off well-defined terminology. Build schema to contain linked data containing values, and annotations on values used to ingest the data. Layered schema approach: meaning is evaluated when look at the data, not when you ingest it. Meaning is context-dependend. May have multiple labels, may ingest with one set, but look at with different set of labels, that is when you interpret. That is what linked data gives you.

Paul: that part need to think of a proper use case. Vaccine work.. Governments to dictate objects. Not creating multidimensional code. For this credential, predefined entries. But want to try to make it so that Burak has enough flexibility to do whatever you want outside those core functions the issuers will need to deliver.

Burak: true: layered schema approach purely as metadata. Well-defined functions, projections, Spar(?), transformations on data. Structured OCA limiting it, in my opinion, greatly, because you are too much focused on the capture part.

Paul: it's my life, not going to argue. Know how important it is. Pharmaceutical. Want to introduce to Linked Data community: work done in Pharma stuff hugely valuable: 100 years of structuring data in a beautiful way. Covid stuff: need to be rigid on the layers. But think can drastically reduce them now. Good: sign of good architecture. Reducing functionality right direction. Like DID stuff. Have 96 different DID methods: they probably got it wrong.

Paul Knowles To Everyone

Yes please, Brent. That would be awesome.

Burak Serdar To Everyone

fyi: Some information about layered schemas is here: https://layeredschemas.org/

Neil: MSFT. Evolution. Good architecture: Core is immutable but extendable. They put display processing in the core and it blew up. That has to be outside the OS - evolves rapidly over time. Some primitives. What is the core? Can add own layers on top. Analytical layer: what rols is each attribute playing. Attributes can play more than one role. E.g. is this attribute indexable? … things to add on top that don't mess with the core.
"Open/Closed Principle." https://en.wikipedia.org/wiki/Open%E2%80%93closed_principle
Can process any way you want, but what places work.

Paul: can come up with something special. Keep decent structure at storage site, but leave maximum flexibility outside that core. From specs now, I know what I need to do at the capture layer - don’t need all the masking. All need ar elabels, predefined entries, diff languages, flag PII. That’s pretty much all need.

Burak: believe the separation isartificial. No diff between capture and storage side. Just open-ended layers with defined terminology. Don’t agree that should have sensitive overlay or label overlay. Just should have terminology. Layers should not have predefined functions, just be layers.

Paul: not how schema definition works. Need structure at capture side.

Burak: not saying get rid of structure, but restructure so it doesn’t have to be separate functional overlays. Still have labels, … enumerations, but they are not separated into different overlays. Just have term in input terminology saying term …, can switch that. As schema-based and overlays.

Paul: defer to Robert

Neil: it’s tomatos/tomatoes. Agree on layered approach. How you slice it. Problem building an ETL model - going rom source A to dest B, need to invent each time the internal schema from which you shuffle all the data through. Suck it in through the model and put it back out. Don’t have to invent interchange piece: If dealing with vac record, don’t need to do that work anymore. … If picking this as core interchange schema, you get all these other things because they have already done that work. … Huge amount of work that just fell off the table.

John Walker: Purpose-built approach. When Gov of China puts out set of code values, there is an intent, the purpose is vaccination receipt/credential. There is a prescriptive set of steps, interpretation of what they publish, that they intend this for its usage. That can be discovered, reassembled, for processing, and in storage. There is a prescription. They are saying there is a context that they are intending the usage for. Reflects back to agriculture which is more prescriptive in pipeline. If know two or three steps or transforms from the source, take advantage of it. Could be discovered, rearranged like Burak saying, yes. But don’t want to forget original language. Keep simple

Burak: agree, but then… the work you did with FHIR, OCA: shows that what you just said didnt’ really work. OCA/FHIR pipeline outputs OCA schemas/artifacts. Not the right way. Should have had OCA do the transformation. Semantic pipeline in the middle to transform. Produce VC using OCA layers. But that’s not how it is working.

John: What we do is start with FHIR bundle, then run a JSON-LD transformation, vs. applying OCA layers, than apply OCA layers after. Think we net to the same place.

For VC/LD Stuff.

Me To Everyone

12:40:15 PM

Anything to say about https://github.com/w3c/lds-wg-charter/ ?

Paul: as semantics team, we should come up with how to go. Naming of stuff not important to me. This whole thing is that my wife doesn't shout at me about that 6G Helsinik app come up , everything digitaized and out landscape not equipt to deal with that level of automation. If we can all come together to a common place for the benefit of the … solution. Arm wrestle in the core. Outside of that, make sure super-easy to use, people do whatever they want. Data super-structured. If throwing data towards WHO, they can pick it up and do analytics in real time to know how the pandemic is responding. I fwe can get there then I'll be a happy boy.

Neil: … where data coming from … Outside VC, data in super-raw form. Same for IOT devices. 5 distinct data organizations to create from the raw stream for different …. Gets bundled into transactions. On top of that, have operational model. Keeps going up, more sophistication in each layer, adding additional metadata. Need to apply some categorization, take a range of values, birthdate in ranges. Lifecycle. Right now all hand-coded. Lived through this in BI stack model - Stack of analytic databases. Things in the real-world, exchanging VCs, overwhelming assumption is we are exchanging the same VC. How to deal with mismatching VCs? What if have composite? Want to look at combination of vaccination dates, whether tested, have recovered. Want to apply country-specific formulat to weight factors to decide whether can come in country.
VC wallet policy execution piece. How do they wwalk tot all these VCs? No mechanism right now. VC exchange thing just be fluffed off - "we're using our VC, of course".

Paul: fascination discussion. Have seen a couple interesting things on data collection side. Linked data stuff, not necessarily what the VC people are using, very sophisticated what the MIT people are doing. Haven't done layered approach - think they need that. Going from that to the rules part. If can do combination of some of that stuff, think it could work. That's as technical as I go.

Neil: Do we have to buy Brent a beer?
Paul: We do, and I have to send you that stuff so you know what I'm talking about.
Tom you should have it, Neil have it, … Is private?
Brent: it's all publicly available.

Paul: plug: Inputs and Sematics WG. Semantics Domain Group Pushing layered architectures. Layer of data collection could be looking at as well. Think will need to take into semantics to understand it.
Every week on a Tuesday.
Interesting group at ToIP. Like an innovation hub. When came into the space, it was like we wheeled in a trojan horse. ToIP initially built by SSI community. Thinking about authentic data, VCs and stuff. When I came in with my data management hat, saying need a semantics group,

fabulous people jumped iout, started building amazing y tealanted group of data magmtnt semantics experts. Got a place at the table. Identity folks recognizing importance of structured data.
…

Neil: plug: watch Buraks breakout video from yesterday.

Neil: Query Intent

Charles: asking about Linked Data Signatures W3C WG?
Neil: Not relevant. Input state. Consumption
John: Interesting work. Best practices encryption. Canonical rep important.

Paul: would want VC to just sign ….

...

Chat

Chat messages may be out of order with regards to the other typed notes.

Neil Thomson, 12:04:14 PM

Machine readable data is not analytically usable without the metadata

...

Scott David, 12:04:57 PM

Data plus meaning equals information. Is semantics co-extensive with meaning?

Neil Thomson, 12:06:21 PM

That's the goal...
Key question - so what metadata is required (and is missing) for data that is useful outside of the original context

Scott David, 12:08:17 PM

Is semantic layer intended to anticipate all possible future contexts of application? If so, how is this open ended future encoded in the system?

Can the semantic layer be Bayesian, so that modifies anticipated contexts and "learns" as time goes on? Is that them semantic active inference?

Linked data concept

Neil Thomson, 12:10:24 PM
Semantics is built from (most cases) several layers.

Brent Shambaugh, 12:15:56 PM
Thanks for summarizing and your work. :)
I believe NULL is a maybe functor?
CQL deals with NULLs . I will need to double check the specifics.
I'd like to participate more if that is useful. I could introduce you to Ryan.

Tom Jones, 12:43:07 PM
I have a different approach to consider

Forget the VC, the user should have their medical records on their phone when they travel anyway. Just go from that HL7 data straight to the VP.

My last meeting was too successful - I'm still trying capture all the data I got from that

Healthcare semantics.

…

Need to know someone they trust authorized a physician
Mixing two problems. Whether data on phone is appropriate. Second question is whether or no have proof presence, of you presenting it. Doesn't need to be part of the same mechanism.
Paul: have separated those. Separated.

Need to cryptographically link back to the source data.

Tom: health data is health data, it is where it is. Huge money spent to make sure its accurate. DOn't ned to do that.
Paul: Still need to cryptographically link to the record. To this credential. If lose that, the credential is meaningless.
Tom: I'm not putting my health in your technology.
Paul: Don't want it, you can keep it.
Tom: Should get data from where it was generated, put in phone and take with you.
Paul?: HL7 is a standard, not *the* standard.
Tom: … Other standards. The data is that way. Where it was captured. Can do that transform. Still will be HL7, but will lose information . No chance to do transform and gain data. Always lose data. Best thing to keep source data, use transformation at end.

Paul: you think hospitals would be happy with that? Real life situation: capture data, goes in their DB, you want to get the same data on your side.
Tom: it’s regularity in force in US today. Must be able to do that.
Paul: layered architetcute. Can have hospital / gov authority to publish overlays… to translate to some sort of VC or proof.
Tom: HL7 is just a structure. It should work everywhere. Just carrying basket. Codes/fields are specific to some regions. Want to keep the data in the format it was generated in, convert it at the last moment. Transformation…
Paul?…
Tom: have to do that contemporaneously.

Paul: have health format, types don’t match. Have to match types. One wants certain types of codes.

John: Conversion could go either RDF to JSON-LD.
Papers for round-tripping data.
FHIR standard is JSON. But also graph representation. Issues
Should be able to go bidirectionly RDF JSON-LD
Limited sense: subset of FHIR resources, can take from graph and put into JSON-LD rep.
Haven’t tried round-tripping it.
FHIR Lab(?)

Burak: Talking about semantic transformations.

John Walker, 1:14:53 PM
https://github.com/fhircat/FHIRCat

Tom: …
Data collection.

John: work as add-on extension. Talking about step beyond that, (transform to LD).
JSON-LD vocabularies.
Transformation to apply to LD resources.

Tom: part of idea insisteninc. You guys are going to stop after having succeeded. Then a new disease and set of codes will be created. If keep data originally in HL7 format, don’t have to recreate it. Have IG do the first transform. In my view, that would be a VP rather than VC.

John: probably requires both. Technique can be generalized to any combination of FHIR resources that make sense. Focus on this use case. Nothing about this technique that is not permutable to any set of valid FHIR resources.

Tom: minor correction: we already have a clinical document for travel. Don't have a public health document. Should be talking about public health document.

John: Event on clinical side.
WHO extending IPS (intl patient summary) - WHOS VC to map/extend from IPS.

Tom: Suggesting have parallel doc for IPS which is already in it for public health. Right start. Take it and modify it. Different set of elements in each one, for privacy reasons and others.

Brent Shambaugh, 1:19:54 PM

> ShEx remind me of Dragon based on C.T. I think Josh mentioned a relation there. Henry Story is also looking at Category Theory. https://web-cats.gitlab.io/

Paul: ^ this is regarding all that categorization/linked-data stuff MIT working on.