**NIST Big Data Public Working Group (NBD-PWG)**

**NBD-PWD-2016/M0511**

**Source: NBD-PWG**

**Status: Draft**

**Title: Web chat from Meeting of March 15, 2016**

**Chat Log D:\\_wo\1DMG\2015\\_BigDataWG\Docs\ChatLog 2016\_03\_15 15\_13.rtf**

**Russell 2 (Private)**: 1:09 PM: WG9 meeting in Dublin last week. CD = committee draft. want to get 20456 into CD. Diff countries vote + comment / feedback. Vocabulary [nist] 3 yr deadline approaching. Other new inputs for RA. Interface framework.

**Russell 2 (Private)**: 1:12 PM: EU PPP planing to spend 2.5 billion on BD initiative. Dave attended.

**Russell 2 (Private)**: 1:15 PM: 20547 was held up. Wanted to split it up, barely passed JTC1 vote but still holding over resolving SnP placement. Looking for contributors. Not clear what areas need contribution. INCITS registration.

**Mark Underwood (to Everyone)**: 1:22 PM: Dave - you mean the two docs dated 10MAR?

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:22 PM: My apologies for being late.....currently slowed down by Bursitis of the right hip......bummer...or as the French would say...."Boom-aire"......

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:22 PM: We'll be quite...Pw

**Lee Anne Davies (to Everyone)**: 1:24 PM: Wo, you have my list of possible contributions I emailed you. Was it helpful?

**Russell 2 (Private)**: 1:26 PM: I need to ramp up my effort to ingest more contributors. NBD prep for Sep workshop is behind schedule.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:26 PM: Dear Ms. Davies: Are you contributions available for download and review.....? pwc.pwcarey@gmail.com......

**David Boyd (to Everyone)**: 1:28 PM: Mark yes

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:29 PM: Dear Ms. Davies: If you're sharp enough to find this WG (Working Group)....that's good enough for me....

**Russell 2 (Private)**: 1:31 PM: May take weeks to get the PWG website back up. Lee Anne: working in Vol 1+2. Lifecycle model, correlation,

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:32 PM: Dear Ms. Davies: Re-No. 4....A lot of C-suite folks follow their gut feelings-hunches, rather than Big Data Analysis....experience, intuition should not be set aside inplace of 'technical-knowledge'......

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:34 PM: Dear Ms. Davies: Nice list....shows alot of thought put into your effort......Good, good & good initial shot....

**Russell 2 (Private)**: 1:35 PM: No response from LA on offer to collaborate on lifecycle. leeannedavies@agenomics.ca

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:36 PM: Re: No. 7----Correlation is wide open for subjective bias...while causation less so.....(aka: the open door caused the data breach....)

**Mark Underwood (to Everyone)**: 1:38 PM: Current discussion: the Point in Time consideration

**Mark Underwood (to Everyone)**: 1:38 PM: classic issue in data warehouses; the big data aspect is probably cross-enterprise

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:39 PM: Public Information: The x Director of NSA stated on an open forum that 'meta-data' was used to remove bad guys from the Terrorist Team side of the field.....the aggregation of meta-data is quite powerful for identification of individuals, with or without a warrant.....Pw

**David Boyd (to Everyone)**: 1:39 PM: Wo all, I need to step out for about 30 minutes.

**Mark Underwood (to Everyone)**: 1:39 PM: Sandra - getting some particulars on that use case could be helpful

**Russell 2 (Private)**: 1:40 PM: Caller 1 = Sandra.

**Russell 2 (to Everyone)**: 1:40 PM: data lake is not a proprietary term or concept

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:41 PM: Dear Mark: "At this point in time"....is easily lost/and or glossed over during sequential comparisions and analysis of data.....our opinion.....

**Russell 2 (to Everyone)**: 1:42 PM: No, no individual brand has adopted the lake term usage.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:43 PM: We agree with Russell....it's a new term to us....

**Russell 2 (to Everyone)**: 1:44 PM: PW - how can you agree with me if the term is new to you?

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:45 PM: Dear Sir: We

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:45 PM: We're just being polite.....go figure....kind Sir....

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:47 PM: & it is a never before heard term by my ears.....go figure....almost like we're not 'current'......

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:48 PM: Dear Tim: Should we also include US Code 645 Entrapment.....?

**Russell 2 (Private)**: 1:53 PM: Sandra asking what level audience she is writing for. Vol 7 serve as high level on how to use rest of the volumes.

**Russell 2 (Private)**: 1:55 PM: Tim saying write it for your supervisor as a multiple choice test.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:55 PM: (Aka: The mis-use and abuse of Big Data....by any and all organizations, individuals and entities)?

**Tim Zimmerlin (to Everyone)**: 1:58 PM: Some people call content management systems (CMSs) data lakes because they can toss anything into the CMS as long as their is a "filter" to convert the format to some base format like XML or PDF.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 1:59 PM: Thanks Tim:....we learned something new, today....

**Tim Zimmerlin (to Everyone)**: 1:59 PM: Sandra's FOIA e-discovery use case would suit a CMS solution, with Solr/Lucene as a good enough too.

**Tim Zimmerlin (to Everyone)**: 2:00 PM: too = tool

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:00 PM: Dear Tim: What's the time-line end-to-end for a FOIA to take effect and the requested data being brought out of the darkened rooms.....?

**Tim Zimmerlin (to Everyone)**: 2:01 PM: Pw, I do not know, but FOIA causes extra work.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:02 PM: Ok...thanks...&...true, true & true....but we believe it's worth the effort from our perspective....Thanks....

**Tim Zimmerlin (to Everyone)**: 2:05 PM: Cavan, you could use URLs, URIs, URNs in linked data for metadata values. Then, Xquery, Sparql, etc.

**Russell 2 (Private)**: 2:05 PM: Cavan: classic metadata was olap on relational. nosql involves diff kinds of data. Ann: predicting what customers will do is a privacy problem.

**Tim Zimmerlin (to Everyone)**: 2:06 PM: Cavan, you could use Solr/Lucene to search for all concrete references to one document, finding all metadata for that single document.

**Lee Anne Davies (to Everyone)**: 2:08 PM: Lee Anne's contact: leeannedavies@agenomics.ca

**Cavan (to Everyone)**: 2:08 PM: Tim, I agree about Solr(or Elastic search), it doesn't construct valid linked data results.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:08 PM: A new security and privacy initiative is taking place in the UK....when will it reach here is open to question....https://www.gov.uk/government/speeches/home-secretary-publication-of-draft-investigatory-powers-bill

**Tim Zimmerlin (to Everyone)**: 2:09 PM: Cavan, I provided 2 independent design patterns there. If XML links cannot work well, then Solr.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:10 PM: https://www.gov.uk/government/speeches/home-secretary-publication-of-draft-investigatory-powers-bill

**Tim Zimmerlin (to Everyone)**: 2:10 PM: Cavan, really old data is scanned & digitized with OCR errors and dropouts, so Solr with spell error distance metrics work.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:11 PM: Well said Ann.....well said....

**Cavan (to Everyone)**: 2:11 PM: pw - thanks

**Russell 2 (Private)**: 2:12 PM: An thinks we need to cautiously approach the knowledge vs. insights dialog, and the stages construct only serves certain purposes and not others.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:12 PM: Your welcome....

**Russell 2 (Private)**: 2:13 PM: Same with the lifecycle [Ann]. I think the lifecycle is not so socially fuzzy.

**Cavan (to Everyone)**: 2:13 PM: Tim, almost 50% of links are wrong, due to collisions with names addresses, misspelling etc.

**Cavan (to Everyone)**: 2:13 PM: so have to use more statistical techniques to get good matchs

**Mark Underwood (to Everyone)**: 2:14 PM: ADP calls their 3-4 petabyte repos of anonymized employee records (24M) a data lake

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:14 PM: Or....clean up the source data....yes...no....?

**Tim Zimmerlin (to Everyone)**: 2:14 PM: Cavan, a curation problem. 80 percent of labor and cost involves data cleaning. Metadata documents cleaning steps / consequences.

**Cavan (to Everyone)**: 2:15 PM: multi-million dollar curation problem... Jim is wrong but James is right etc.

**Tim Zimmerlin (to Everyone)**: 2:16 PM: Cavan, effective use of metadata includes reification of entities / identities. E.g., John Smith 123.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:16 PM: Soooo, if there is no Cost Benefit Analysis justification and/or a defensible one....then the errors remain in-place....no?

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:17 PM: As in, it's not worth the effort to be correct....

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:18 PM: That is not until a severe enough pain point is reached...then the correction will be maide....our opinion...

**Cavan (Private)**: 2:21 PM: wo - sorry if I was slow on the data lake

**Russell 2 (Private)**: 2:22 PM: Tim: value of data exists in kidney marketplaces, specialized job databases, aggs who resell, value question is hard to answer.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:24 PM: Dear Ms. Davies: Data Horders vs Data Tossers.....(aka: human nature kicks in.....which cuts both ways....for good or ill....) our opinion...

**Russell 2 (Private)**: 2:27 PM: LA: hoarding is indiv; glut is organ12l

**Tim Zimmerlin (to Everyone)**: 2:32 PM: FCC has just announced they will make the policy for privacy for all Internet Service Providers. Already, a big storm of criticism from industry...

**Russell 2 (Private)**: 2:32 PM: Cavan: consent is evolving. Boundaries not clear.

**Tim Zimmerlin (to Everyone)**: 2:34 PM: Cavan, the big privacy sticks are insurance / liability and lawsuits and criminal prosecution.

**Russell 2 (Private)**: 2:36 PM: 60% of IBMs business is outside the US. No strategic doc about how to make privacy viable.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:36 PM: Dear Tim: Should we question state sponsored abuse.....maybe....?

**Cavan (to Everyone)**: 2:37 PM: Tim, yes and that is a continuing process. My problem is that it is often ad hoc, and there seems to be no strategic arch view to facilitate

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:38 PM: Take it easy...nice to hear you voiice and thoughs....too....

**Russell 2 (Private)**: 2:38 PM: LA: access ctrl only at the \_ level. her Vol 4 section 3b. Somet beyond a technology such as ethics.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:39 PM: that would be your voice and thoughts....Ms. Davies...

**Cavan (to Everyone)**: 2:39 PM: A seminar really on privacy threats due to record linkage

**Cavan (to Everyone)**: 2:39 PM: no data can be private due to outside datasets, without formal differential privacy being applied

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:40 PM: and maintained.....

**Tim Zimmerlin (to Everyone)**: 2:40 PM: Cavan, conclusion understood.

**Cavan (to Everyone)**: 2:40 PM: differential privacy puts noise in the data in such a way that the stats are the same but individual values are different

**Cavan (to Everyone)**: 2:41 PM: .... and value of data is no longer just a ethic.. a differentially private data set will have a privacy budget associated with it

**Tim Zimmerlin (to Everyone)**: 2:42 PM: Cavan, today, scientific / statistical reproducibility are very important…especially in social sciences.

**Cavan (to Everyone)**: 2:42 PM: and certain fields will be perturbed more than others

**Cavan (to Everyone)**: 2:43 PM: yes, I have worked closely with the Harvard Dataverse which is a leader in reproducibility

**Cavan (to Everyone)**: 2:44 PM: but they are also a leader in differential privacy, along with Cynthia Dword at Microsoft

**Cavan (to Everyone)**: 2:45 PM: so if you are transferring data across the enterprise without making it private, that is an IT security issue, not a privacy issue

**Tim Zimmerlin (to Everyone)**: 2:45 PM: Hey, Yuri! Welcome back.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:46 PM: All individual privacy should be a manual opt-in process....otherwise...the individual is identified via an anonymous construct....only one direction...either you agree to be included or you're not included in the dataset.....simple, no?

**Cavan (to Everyone)**: 2:47 PM: PW - I wish, Denmark had opt out policy on health/medical records and had an approximate 80% participation, they changed it to an opt in.. and participation fell to around 26%

**Arnab Roy (to Everyone)**: 2:48 PM: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded\_Top\_Ten\_Big\_Data\_Security\_and\_Privacy\_Challenges.pdf

**Mark Underwood (to Everyone)**: 2:48 PM: I have to leave the call- thanks all

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:49 PM: Dear Mark: Have a nice & Happy Saint Patrick's Day....

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:49 PM: Dear Cavan: So the folks in Denmark wanted to remain anonymous rather than Opt. IN....or do we have this backwards....?

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:50 PM: They wanted to keep their medial records private....?

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:51 PM: Medical

**Cavan (to Everyone)**: 2:51 PM: Denmark has been a leader in using administrative records available for statistical/medical research

**Russell 2 (Private)**: 2:51 PM: demch = Yuri \_.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:51 PM: But is the privacy of the individual maintained....?

**Cavan (to Everyone)**: 2:52 PM: the government changed and changed the opt-out process for allowing medical records for research to an opt-in

**Cavan (to Everyone)**: 2:52 PM: and research significantly declined

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:53 PM: So, you had to manually join in in the research otherwise you were not included....ok that makes sense to us now....Thanks....

**Bob Marcus (to Everyone)**: 2:53 PM: 15 slide sets on IoT, Cloud, and Big Data at http://www.slideshare.net/bobmarcus/inventory-of-my-cps-slide-sets

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:55 PM: The EU European Union takes a different view of individual's identity and privacy than in the United States of America....at this point in time....

**Cavan (to Everyone)**: 2:55 PM: agreed, and each country has variations (i.e. Germany)

**Russell 2 (Private)**: 2:55 PM: Wo: how to address IoT aspects. Bob: moves us toward relevancy; where big data is now already dealt with / less uncertainty. IoT draw increasing participation; should have a joint meeting [cloud, bd, iot].

**Russell 2 (Private)**: 2:56 PM: in Sept.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:56 PM: & the Irish just don't care.......

**Cavan (to Everyone)**: 2:56 PM: all that stout

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 2:56 PM: You are CORRECT Sir.....

**Russell 2 (Private)**: 2:58 PM: Uncertainty, unfinished work by any authors = opportunity.

**Ann Racuya-Robbins (to Everyone)**: 2:59 PM: I have to drop thank you.

**Tim Zimmerlin (to Everyone)**: 2:59 PM: Bob, good talking points. Most folks agree with you. Easier said than done, though.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 3:00 PM: Dear Sir: Once again you have forced upon us a very useful and constructive meeting.....must leave to attend an early afternoon Saint Patrick's Day Festival....Thank you....Respectfully yours, Pw

**David Boyd (to Everyone)**: 3:01 PM: Got to leave. Another meeting.

**Russell 2 (Private)**: 3:03 PM: William: transactive energy use case. Setup call for Mark to 'interview' for the new template. EU standards bodies are more active than we are. 1mtm. Consolidation of consortium outside the traditional ieee; formed a way of sharing data. PIA and OMA i DC are participating. Detailed docs.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 3:05 PM: Dear William: Is there a link to those organizations in the EU....?

**Russell 2 (Private)**: 3:06 PM: Not mention meta that propagates w data. EU concerned about the US collecting data on their citizens.

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 3:07 PM: Dear Sir: Thank you.......

**Pw Carey, Compliance Partners, LLC (to Everyone)**: 3:08 PM: Yuri-dim-chee-cough.....nice name, Yuri....

**demch (to Everyone)**: 3:08 PM: demch = Yuri Demchenko

**demch (to Everyone)**: 3:09 PM: my audio dropped

**Russell 2 (Private)**: 3:10 PM: Yuri Demchikov. Amsterdam.

**Tim Zimmerlin (to Everyone)**: 3:12 PM: Yuri, I nominate you to be the EU Commision representative to NBD-PWG.