Recap: twc-healthdata, TWC's submission to Department of Health and Human Services' Developer Challenge

Jim McCusker, Timothy Lebo, Alvaro Graves

TWed@TWC
23 Jan 2013

Talk homepage: https://github.com/jimmccusker/twc-healthdata/wiki/2013-Jan-23-twed

Outline

- What they asked for / What we did
- Demo
- Lessons Learned
- Future work
  - Wish we could-a
  - Some re-thinks
  - Follow-ons inspired by the HHS challenge
- Conclusions

What the *Department of Health and Human Services* Asked For

"...establish learning communities that collaboratively evolve and mature the utility and usability of a broad range of health and human service data."

HealthData.gov

-http://www.healthdata.gov/developer-challenges-overview

385 dataset listings at

http://hub.healthdata.gov

What they wanted

- Metadata
- Classification
- Liquidity
- Automation
- Documentation
- Engagement

They wanted *Metadata* part 1

"... application of existing voluntary consensus standards for metadata common to all open government data"

We gave them

- **DCAT - W3C Data Catalog**
  - Version controlled on github.
  - Extracted from their CKAN as input to converter.
- **VoID - W3C Vocabulary of Interlinked Data**
  - Organizes datasets by source, dataset, version.
  - Provides links to data dumps, Linksets to LOD.
- **PROV - W3C Provenance Interchange Model**
  - Captured during CKAN extraction, retrieval, conversion, and publishing.
- **Dublin Core Metadata Terms**
  - Annotated subjects based on descriptions.

They wanted *Metadata* part 2

"Metadata tags that have [dereferencable] HTTP URIs..."

We gave them

- Converter creates URIs for all data values
- LODSPeaKr publishes SPARQL endpoint as Linked Data
- Reused entity URIs from
  - LOGD's Instance Hub (states, providers)
  - Bioportal: ICD, SNOMED-CT (hospital types)
- Conversion metadata tailored to the dataset
- Pinging ***http://sindice.com*** at each dataset update
- Refreshing ***http://datahub.io/dataset/twc-healthdata***

They wanted *Classification*

"...classify datasets in our growing catalog, creating entities, attributes and relations that form the foundations for better discovery, integration..."

We gave them

- Bottom-up vocabulary and entity reuse
    - Vocabulary created for each dataset
    - Enhanced datasets shifted to reuse vocabulary and entities from other datasets.
    - Three stub vocabularies for top-level reuse.
- NCBO (Nat. Center for Biomedical Ont.) Annotations
    - annotator/annotator.py SADI service
    - data/source/bioontology-org/annotator-description-subject/version/retrieve.sh

They wanted ***Liquidity***

"new designs ... that form the foundations for ... liquidity"

We gave them: 2B triples among 1M URIs

- Dataset Linked Data
  - Machine and Human views (via conneg)
  - Faceted search of datasets
- Dataset dumps (.ttl.gz)
  - For each dataset, and for *the whole thing.*
- Dataset query (http://healthdata.tw.rpi.edu/sparql)

They wanted *Automation*

"Tools: use of automation"

We applied:

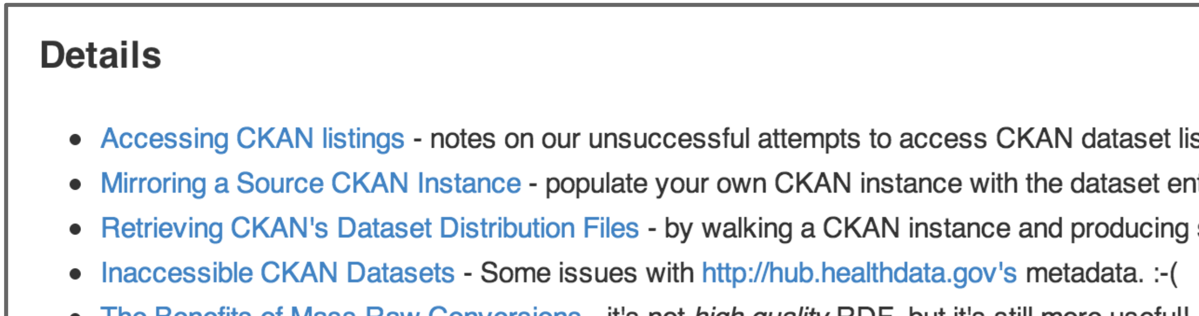
- *https://github.com/jimmccusker/**twc-healthdata**/wiki*
  - Version-controlled essential bits.
  - Provides basis for anybody to recreate what we did.
  - Provides infrastructure for anybody to contribute.
  - Forkable - no need to coordinate or get permission.
  - Nightly cron - (idempotency!).
- *https://github.com/timrdf/**csv2rdf4lod**-automation/wiki*
  - Automatically catalog, retrieve, convert, and publish.
  - Driven by RDF metadata about the datasets.
- *http://alangrafu.github.com/**lodspeakr***
  - Publishes 5-star Linked Data.
  - Provides aggregated data (list of datasets).

They wanted *Documentation*

"Documentation: articulation of design using well known architecture artifacts."

We gave them:

- https://github.com/jimmccusker/twc-healthdata/wiki
  - 19 wiki pages describing each stage/component.
  - + links to original documentation for each tool that we used in each solution.
  - Written *while* the work was done; shared among the collaborators as status updates.



Details

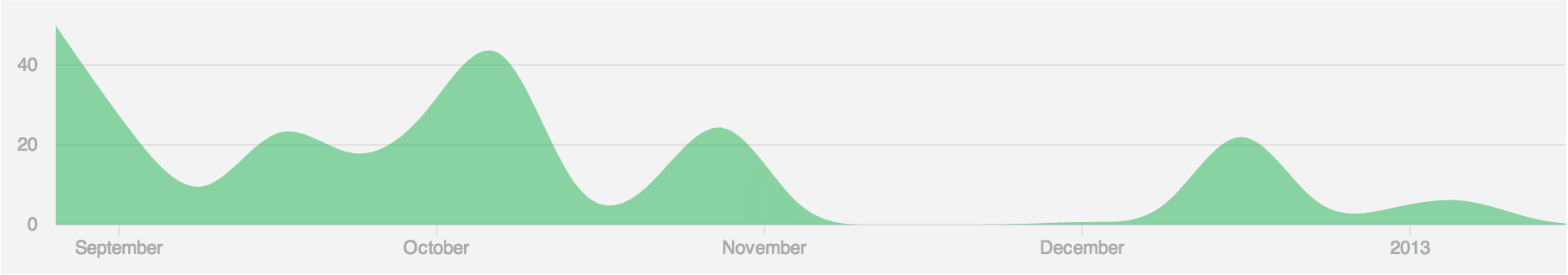
- Accessing CKAN listings - notes on our unsuccessful attempts to access CKAN dataset lis
- Mirroring a Source CKAN Instance - populate your own CKAN instance with the dataset en
- Retrieving CKAN's Dataset Distribution Files - by walking a CKAN instance and producing
- Inaccessible CKAN Datasets - Some issues with http://hub.healthdata.gov's metadata. :-(
- The Benefits of Mass Raw Conversions - it's not *high quality* RDF, but it's still more useful!

They wanted *Engagement*

"Engagement: willingness to participate in the community as a maintainer/committer after award"

We gave them:

- https://github.com/jimmccusker/twc-healthdata
    - We're still going!
- twc-healthdata benefits from ongoing developments for subsequent applications (with a just a git pull)



Demo

http://healthdata.tw.rpi.edu

https://github.com/jimmccusker/twc-healthdata/wiki

http://healthdata.tw.rpi.edu/hub/

Lessons Learned (Technology)

- Version control!
  - ***github***
  - Develop from a writable working copy
  - Publish from a read-only working copy
  - Publish from a project-specific user name
- *Everything* is a Versioned Dataset.
  - (even cron)
- Don't name RDF files after their vocab use.
  - e.g. "void.ttl" should have been "meta.ttl"
  - e.g. "pml.ttl" should have been "prov.ttl"
  - and we still went with "dcat.ttl" :-/
- It's tough to training people to model well.
  - "Good" needs a grounded, realistic definition.

Future Work:
Wish We Could-a

- ... had better access metadata from HHS.
- ... did more comprehensive raw-value analysis to recommend dataset curations.
- ... provided better navigation of the vocabularies used
- ... created better transition between data and vocabulary.

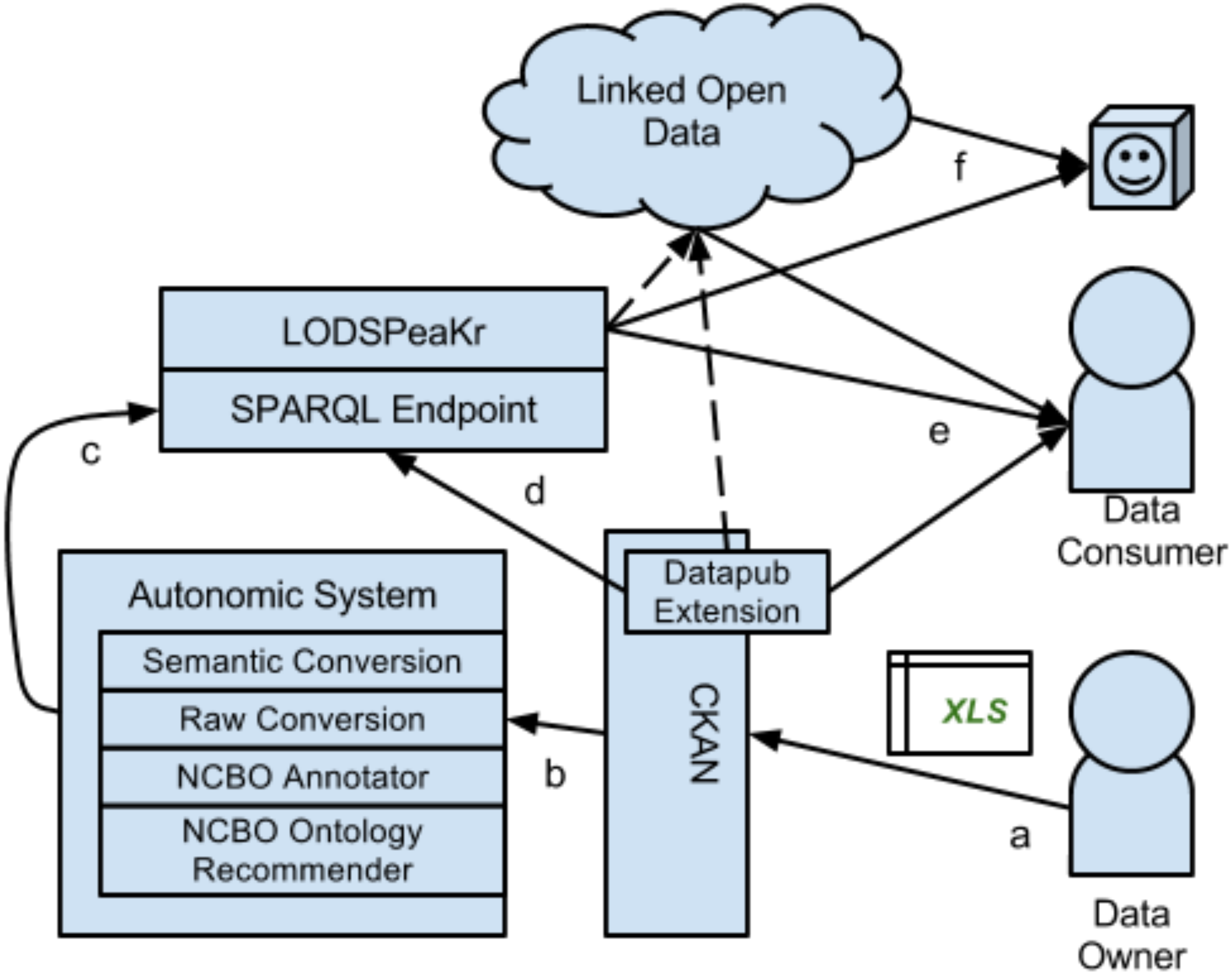
Future Work:
Some Re-Thinks

- healthdata.tw.rpi.edu vs. purl.org deathmatch
- Organizing datasets by *retrieval source* (instead of catalog provider)
    - e.g. data-gov when the file is from epa-gov
- Mix-and-matching LODSPeaKrs
    - Document decentralized approach of development
- Provide better debugging tools for LODSPeaKr
- Redefine DCAT's "accessURL" -- too ambiguous!
    - *"...can be a direct download link, a link to an HTML page containing a link to the actual data, Feed, Web Service etc."*

Future Work:
Follow-ons inspired by twc-healthdata

- Prizms - Better Visualizations Catalyzed by Better Data
  - csv2rdf4lod-automation + DataFAQs + LODSPeaKr
- SPO Balance
  - Vocabulary: http://prefix.cc/vsr
  - Overview+detail for any RDF dataset
  - Sesame implementation produces summary descriptions of a triple store using DCAT + VoID + SD + SIO + PROV
  - 500+ triples to describe TBL's 79 triple FOAF file :-)
- Between The Edges - Explicit semantics of single URIs
  - Vocabulary: http://prefix.cc/bte
  - Implementation: sadi/faqt/naming/between-the-edges.py
  - 379 of (1M healthdata.tw URIs) -> 3,288 triples (1B triples?)
- lod.melagrid.org - Applying Prizms to melanoma data.

Looking Forward: The Next Data Sharing Architecture



Conclusions

- Demonstrated Linked Data for HealthData.gov Platform

- Collaborating to build a system is easier to do when the parts connect using the semantic web
    - Especially when in a volunteer, *ad hoc* environment

- Didn't just make health data *better*, it made future Linked Data *better*!
    - Inspired a flood of new features for our tools
    - Uncovered a handful of bugs (that we fixed ;)

- Still plenty to do!

Backup: Vocabulary use

DCAT
(7/19 terms)

<http://www.w3.org/ns/dcat#Dataset>	12,845
<http://www.w3.org/ns/dcat#Distribution>	3,968
<http://www.w3.org/ns/dcat#accessURL>	3,958
<http://www.w3.org/ns/dcat#dataDictionary>	123
<http://www.w3.org/ns/dcat#distribution>	3,968
<http://www.w3.org/ns/dcat#granularity>	314
<http://www.w3.org/ns/dcat#keyword>	15,246

VoID
(11/29 terms)

void:Dataset	7,220
void:DatasetDescription	73
void:Linkset	15
void:dataDump	1,172
void:exampleResource	2,202
void:inDataset	1,238,891
void:rootResource	3
void:subset	3,465
void:target	6
void:triples	1,252
void:vocabulary	7,178

PROV (19/80 terms)

prov:Activity	85,182
prov:Association	77,249
prov:Plan	10
prov:agent	487
prov:alternateOf	6,633
prov:atLocation	7055
prov:endedAtTime	836
prov:generatedAtTime	8688
prov:hadPlan	83,628
prov:qualifiedAssociation	83,287
prov:specializationOf	28,443
prov:startedAtTime	7,664
prov:used	35,902
prov:value	314
prov:wasAssociatedWith	487
prov:wasAttributedTo	31,6605
prov:wasDerivedFrom	9,873
prov:wasGeneratedBy	84,238
prov:wasInformedBy	6,272

DC Terms (23/75 terms)

<http://purl.org/dc/terms/Agent>	6
<http://purl.org/dc/terms/AgentClass>	1
<http://purl.org/dc/terms/contributor>	2,074
<http://purl.org/dc/terms/created>	4,887
<http://purl.org/dc/terms/creator>	2,375
<http://purl.org/dc/terms/date>	3,248
<http://purl.org/dc/terms/description>	22,858
<http://purl.org/dc/terms/extent>	518
<http://purl.org/dc/terms/format>	7489
<http://purl.org/dc/terms/hasPart>	2,077
<http://purl.org/dc/terms/hasVersion>	98
<http://purl.org/dc/terms/identifier>	20,161
<http://purl.org/dc/terms/isPartOf>	317
<http://purl.org/dc/terms/isReferencedBy>	350,231
<http://purl.org/dc/terms/issued>	659
<http://purl.org/dc/terms/license>	1
<http://purl.org/dc/terms/modified>	10,489
<http://purl.org/dc/terms/publisher>	3
<http://purl.org/dc/terms/relation>	12,794
<http://purl.org/dc/terms/source>	984
<http://purl.org/dc/terms/subject>	2,853
<http://purl.org/dc/terms/title>	1,931
<http://purl.org/dc/terms/type>	2