SEMANTIC WEB IN THE PHARMACEUTICAL INDUSTRY
A LANDSCAPE OF CHALLENGES AND OPPORTUNITIES

- Tim Williams

SWAT4HCLS
Antwerp, Belgium
2018-12-03
OUTLINE

1. Introduction
2. Data
3. PhUSE
4. The Way Forward
5. General Discussion
OUTLINE

1. Introduction
   2. Data
   3. PhUSE
   4. The Way Forward
   5. General Discussion
WHO AM I?

I AM NOT THE EXPERT
YOU ARE LOOKING FOR

Interactive!
Questions and Discussion
WHO I AM

UCB BioSciences
- Statistical Systems Analyst
- Raleigh, North Carolina

PhUSE
- Steering Committee: "PhUSE Computational Sciences Symposium"
- Co-lead: "Clinical Trials Data as RDF"*
- Co-lead: "Analysis Results Model (RDF Data Cubes)" (2016)
- Instructor: "Linked Data Hands-on Workshop"*
I ALSO LIKE #LINKEDDATA MEMES

@NovasTaylor
WHO ARE YOU?

Hands up:

- Pharmaceuticals (any size pharma)
- Biotechnology (non-pharma)
- Health Care
- Research
- Other
WHO ARE YOU : SEM WEB ADOPTION?

Hands up if you are:

- Doing something (personally, professionally) with Semantic Web

Keep your hands up if you are using SW at work in:

- any way: Experiment, Prototype, Proof of Concept, Pilot, Production
- in a Validated Production Environment
OUTLINE

1. Introduction
2. Data
   2.1 Landscape
   2.2 Standards
3. PhUSE
4. The Way Forward
5. General Discussion
2.1 DATA LANDSCAPE

Non-clinical (Pre-clinical)

• Animal studies
## Clinical
- Human Study Subjects

<table>
<thead>
<tr>
<th>Phase</th>
<th>n</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>~15</td>
<td>Safety</td>
</tr>
<tr>
<td>I</td>
<td>~20 - 80</td>
<td>Safety, Dosing</td>
</tr>
<tr>
<td>II</td>
<td>~100's</td>
<td>Safety, Treat Condition, Refine methods</td>
</tr>
<tr>
<td>III</td>
<td>~3,000</td>
<td>Efficacy, Double-blind. Comp. other treatments.</td>
</tr>
<tr>
<td>IV</td>
<td>1000's</td>
<td>Post-approval. Long term efficacy, safety...</td>
</tr>
</tbody>
</table>
DATA SOURCES

Traditional
- Case Report Forms (CRF)
- Electronic Data Capture (EDC)
- * Relational Database Management Systems (RDBMS)

New
- Wearables, Ingestibles, Devices
- Social Media
- Real World Evidence
  - See: openEHR - The 'open platform' Revolution
    Room A, 17:00-18:00

Other Data Sources?
DATA SOURCES (RDF)
RDF ENDPOINTS FOR LATE PHASE DATA?

https://old.datahub.io/dataset/linkedct

Your Experience?
LinkedCT

Data source: Linked Clinical Trials
Size of download and data set: 23 million triples as of April 2011. 4.300 N-Triples dump

CC by-snc license

You are free to copy, distribute, transmit, and adapt the work for non-commercial purposes. However, you need to follow the terms and conditions available at http://www.clinicaltrials.gov/ct2/info/terms

According to O2D (http://www.open2india.org/o2d) the data is not considered open.

Data and Resources

Download

SPARQL endpoint

Data Browser Interface

Example resource

Creative Commons Non-Commercial (NC)

Additional info

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td><a href="http://linkedct.org">http://linkedct.org</a></td>
</tr>
<tr>
<td>Author</td>
<td><a href="mailto:linkedct@helsinki.fi">linkedct@helsinki.fi</a></td>
</tr>
<tr>
<td>Maintainer</td>
<td><a href="mailto:linkedct@helsinki.fi">linkedct@helsinki.fi</a></td>
</tr>
<tr>
<td>Last Updated</td>
<td>July 30, 2015, 09:30 AM (UTC+4:23)</td>
</tr>
</tbody>
</table>
2.2 STANDARDS

HEALTH LEVEL 7

"A set of international standards for transfer of clinical and administrative data between software applications used by various healthcare providers."

HL7® FHIR®
"A draft standard describing data formats and elements and an application programming interface for exchanging electronic health records. Created by Health Level Seven."

FHIR as RDF
Who is using FHIR?
Who is using FHIR as RDF?

Who is attending:
HL7 FHIR and the Semantic Web
Harold Solbrig
Room A, 13:30
Are you using CDISC as RDF?

Are you using CDISC?

If you are in Pharma and not using CDISC Standards, I am worried about you.
CDISC STANDARDS ARE A GOOD THING
BUT THERE ARE PROBLEMS
AND CHANGE IS NEEDED
SDTM DOMAINS

- Demographics (DM)
- Vital Signs (VS)
- Adverse Events (AE)
- ...

"23 defined domains within six broad categories." (SDTM 3.1)
PROBLEMS IN CDISC SDTM

"Domains represent discrete categories" - CDISC

Reality Check: They do not.

- Example: Demographics Domain (DM)
  Also contains
  - Study ID
  - Treatment Arm Information (arm, coded value for arm)
  - Age units
PROBLEMS IN CDISC SDTM

• Multiple approaches to represent Medical conditions
  ▪ Medical History (MH)
  ▪ Adverse Events (AE)
  ▪ Clinical Events (CE)

• Multiple locations for same/similar information
  Death Information:
  ▪ Demographics (DM)
  ▪ Disposition (DS)
  ▪ Adverse Events (AE)

• …and more.
PROBLEMS IN CDISC SDTM

- Data Repetition
- Row-by-Column Structure
THE VERSIONING PROBLEM

- Standards Change over time
- Version-Conversion
  - Instance data is not version-independent
A TECHNOLOGY TIMELINE: XPT FORMAT FOR FDA SUBMISSIONS
IT GETS WORSE...
THE 30 YEAR-OLD XPT FORMAT FOR FILE TRANSFER...

IS BEING USED AS A STRUCTURE FOR DATA STORAGE
Have you converted data from one CDISC SDTM or ADaM version to another?
How can we replace XPT files?
CDISC IS TRYING TO CHANGE
CDISC PROOF OF CONCEPT

"Evolving our standards towards end to end automation"

Image courtesy of P. van Reusel, Sam Hume. "CDISC Proof of Concept: Evolving our standards toward end to end automation."
Example of biomedical concept

- Data is expressed through the CDISC Foundational Models
- Can map to BRIDG Reference Model

Image courtesy of P. van Reusel, Sam Hume. "CDISC Proof of Concept: Evolving our standards toward end to end automation."
Will CDISC Succeed?
3. PhUSE

3.1 What is PhUSE?
3.2 PhUSE Linked Data Initiatives
3.3 CTDasRDF Project
   • The Way Forward
   • General Discussion
Pharmaceutical Users Software Exchange

Mission:
Provide an welcoming, **neutral platform** for creating and sharing ideas, implementing data standards, processes, and tools, and exploring **innovative methodologies, techniques, and technologies**.
Pharmaceutical Users Software Exchange

**Working Groups Mission:**

Provide an open, transparent, and **collaborative** forum in an **non-competitive** environment in which Regulators, Life Science Companies, Technology Providers, SDOs, and Academia can address unmet computational science needs impacting product development and regulatory review as to improve human health

* - emphasis is mine
Pharmaceutical Users Software Exchange

- Membership: >8,700 spanning 30 countries
- Annual Conference: EUConnect, USConnect
- Single Day Events
- Computational Sciences Symposium (CSS)
  - A "working" conference
HANDS-ON WORKSHOP: 21 MERGED STUDIES
PHUSE SEMANTIC WEB (LINKED DATA) PROJECTS

Completed:

- CDISC Foundational Standards in RDF
- CDISC Conformance Checks (incomplete? Last update 2014?)
- Reusing Medical Summaries for Enabling Clinical Research [Demo, P.O.C]
- Analysis Results and Metadata (2016) [P.O.C]
PHUSE SEMANTIC WEB (LINKED DATA) PROJECTS

Past

• Regulatory Guidance in RDF *(incomplete?)*
• Clinical Program Design in RDF *(incomplete?)*
• CDISC Protocol Representation Model in RDF *(on hold [indefinitely?]*)

Current

• Clinical Trials Data as RDF
• Understanding RDF/Linked Data for Nonclinical Use [NEW]
OBSERVATION:
CDISC AND PHUSE PROJECTS HAVE (MOSTLY) BEEN MODELING THE DATA STANDARDS

What is fundamental problem with this approach?
It does not model the clinical trial data.

Proposal:
• Model the Clinical Trial process and instance data
• Build the standards, data checks, etc. - into that model
• Instance data independent from Industry Standards
  ▪ Materialize instance data into a Standard
1. Introduction
2. Data
3. PhUSE

4. The Way Forward
   4.1 Roofshot Manifesto
      4.1.1 Roofshot: Study URI
      4.1.2 Roofshot: SDTM Domains as RDF
      4.1.3 Roofshot: Open Source Ontology Development

5. General Discussion
Linked Data Business Case

Image Attribution: https://bit.ly/2x0Hjmd
4.1 THE ROOFSHOT / MOONSHOT MANIFESTO

**Roofshot**
Incremental impacts
- Study URI
- CTD (SDTM) as RDF
- Open Ontology Development

**Moonshot**
Invent & apply state-of-the-art Knowledge Graph
Clinical trial lifecycle

Concept & Image Attribution: https://rework.withgoogle.com/blog/the-roofshot-manifesto/
4.1.1 *ROOFSHOT*: STUDY URI AS AN INDUSTRY STANDARD (PROPOSAL)


...and additional followup by J. Ulander (A3), T. Williams (UCB)

**Why?**
- Easy entrypoint for Pharma
- Familiar concept: NCT ID
  - CT.gov must first review and approve Protocol
STUDY URI COMPONENTS

https://data.pharma.abc/clinicaltrial/D3562C00096

1. Global Namespace
2. Resource type
3. Trial ID

Is anyone using a Study URI/IRI?
STUDY URI: GLOBAL NAMESPACE

https://data.pharma.abc/clinicaltrial/D3562C00096

- Company web URL
- URIs that de-reference: External/Internal

Discuss
STUDY URI: RESOURCE TYPE
https://data.pharma.abc/clinicaltrial/D3562C00096

• Easy? What else could it be called?
• Implications? Link to ontology?

Discuss?
STUDY URI: TRIAL ID

https://data.pharma.abc/clinicaltrial/D3562C00096

1. NCT ID available (ClinicalTrials.gov)
2. NCT ID not available: Unique Company ID (Company guidance)

Discuss
STUDY URI: NEXT STEPS

- Review and comment at: https://github.com/phuse-org/LinkedDataEducation/blob/master/doc/StudyURI.md
- Invite comment from FDA, EMA, PMDA, CDISC... *You!*
4.1.2 **ROOFSHOT**: CTD (SDTM) AS RDF

PHUSE PROJECT: CLINICAL TRIALS DATA AS RDF
CTD AS RDF  PROJECT PHILOSOPHY

DO NOT MODEL:
- Industry Standards

DO MODEL:
- Clinical trial process
- Data
- Rules
IMPLICATIONS "UP STREAM" FROM CLINICAL STUDIES

PHUSE PRE-CLINICAL JOINS CLINICAL RDF PROJECT!

Common Concepts: Pre-Clinical & Clinical Research
# CDISC Standards in the Clinical Research Process

<table>
<thead>
<tr>
<th>Non-clinical</th>
<th>Clinical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organize</td>
<td>Plan</td>
</tr>
<tr>
<td>SEND</td>
<td>PRM</td>
</tr>
</tbody>
</table>

**SEND**
- Tabulation for Animal Studies

**PRM**
- Model for Planning

**CDASH**
- Model for Data Collection

**SDTM**
- Implementation Guide for Tabulation
- Implementation Guide for Medical Devices
- Implementation Guide for Pharmacogenomics
- Implementation Guide for Associated Persons
- Questionnaires, Ratings and Scales
- Model for Tabulations of Study Data

**ADaM**
- Analysis Data Model

**Data Exchange**
- ODM-XML: Operational Data Model
- Define-XML: Dataset Metadata
- Dataset-XML: Dataset Data
- CTR-XML: Clinical Trial Registry
- RDF: Resource Description Framework
LEVERAGE EXISTING ONTOLOGIES

When you try to choose ontologies for your Knowledge Graph
LEVERAGE EXISTING ONTOLOGIES

BIOMEDICAL RESEARCH INTEGRATED DOMAIN GROUP MODEL (BRIDG)

Collaboration:
- CDISC, HL7, NCI, caBIG, FDA
- OWL version from NCI
- Version 3.2 as RDF. Current is 5.x?
LEVERAGE EXISTING STANDARDS

SDTM Terminology

**study**

"mini" study Ontology

SNOMED
Systematized Nomenclature of Medicine
*Indication Condition*

MedDRA
Med. Dictionary for Regulatory Activities
*Adverse Events*

WHO Drug Dictionary
World Health Org. Drug Dictionary
*Medical Products*

ND-FRT
National Drug File Ref. Terminology
*Pharmacologic Class*

UNII
Unique Ingredient Identifier
*Substances*

RDF

non-RDF
4.1.3 ROOFSHOT: OPEN SOURCE ONTOLOGY DEVELOPMENT

Can an individual developer, project team, company, standards org., or regulatory org. create a solution for the industry?
"We cannot compete with centralized systems unless we collaborate."

- Ruben Verborgh, Decentralizing the Semantic Web Through Incentivized Collaboration
OPEN SOURCE MODEL FOR CLINICAL TRIAL ONTOLOGIES DEVELOPMENT

- Ontologies on GitHub?
- Cooperation in the pre-competitive space
  - PhUSE?
  - TransCelerate?
    - Common Protocol Template (not in RDF!)
  - CDISC?

Discuss
OPEN SOURCE ONTOLOGY

CHALLENGES

• Gate keeper?
• Will companies:
  ▪ Participate?
  ▪ Give back?
• Conflict resolution (approach, code)
• Volunteers

Is this feasible?
ONTOLOGY MAINTENANCE AND DISTRIBUTION

Will the Open Biological and Biomedical Ontology (OBO) approach work?

The OBO Foundry

Post-development curation?
ONTOGRAPHY MAINTENANCE AND DISTRIBUTION

Don't hide my OWL behind an API!
OUTLINE 5

1. Introduction
2. Data
3. PhUSE
4. The Way Forward

5. General Discussion
ADDITIONAL DISCUSSION POINTS
DISCUSSION:

How are we hindering our own progress?

- "High Priesthood"
- Too much emphasis on "Linking all the things?"
- Poor communication, translation to ROI?

What are we doing right?
DISCUSSION:

What are our main challenges in Pharma?

- Momentum of legacy technology
- Skill set, lack of knowledge
- Politics: Who owns innovation?
  - IT
  - Analytics
  - The "Business"
DISCUSSION:

Which is the best environment for SW in BioTech/Pharma?

- Startups, Small Pharma
- Mid-Sized
- Large Pharma
DISCUSSION:

What are you using for validation (& why?)

• SPIN
• SHEX
• SHACL
• OTHER?
What are you using for visualization?

- Commercial Applications
- Open Source Tools
- Bespoke
  - Python
  - Javascript (D3JS, other?)
  - R, RShiny
  - Other?
How do we make the Semantic Web happen in Pharma?
Use Linked Data!!
Thank you!