PHDD – An RDF Vocabulary for the Physical Data Description

*Work in Progress*

Joachim Wackerow and Thomas Bosch
(both GESIS – Leibniz Institute for the Social Sciences)
What is it?

• Description of the physical properties of a data file.

• Focus on most common format types
  – Rectangular format
  – Character-separated values (CSV) or fixed-record length
<table>
<thead>
<tr>
<th>State_FIPS_Code</th>
<th>County_FIPS_Code</th>
<th>CHSI_County_Name</th>
<th>CHSI_State_Name</th>
<th>CHSI_State_Abrbr</th>
<th>State</th>
<th>County</th>
<th>Population</th>
<th>Poverty</th>
<th>Age</th>
<th>Population</th>
<th>Poverty</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>01,001</td>
<td>Autauga,Alabama,AL,29</td>
<td>&quot;frontier status, population size, poverty, age&quot;</td>
<td>37,48612</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,003</td>
<td>Baldwin,Alabama,AL,16</td>
<td>&quot;frontier status, population size, poverty, age&quot;</td>
<td>27,16258</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,005</td>
<td>Barbour,Alabama,AL,51</td>
<td>&quot;frontier status, population size, poverty, age, populati</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,007</td>
<td>Bibb,Alabama,AL,42</td>
<td>&quot;frontier status, population size, poverty, age&quot;</td>
<td>53,21516,81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,009</td>
<td>Blount,Alabama,AL,28</td>
<td>&quot;frontier status, population size, poverty, age&quot;</td>
<td>39,55725</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,011</td>
<td>Bullock,Alabama,AL,75</td>
<td>&quot;frontier status, population size, poverty, age, populati</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,013</td>
<td>Butler,Alabama,AL,76</td>
<td>&quot;frontier status, population size, poverty, age, populatic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,015</td>
<td>Calhoun,Alabama,AL,6</td>
<td>&quot;frontier status, population size, poverty&quot;</td>
<td>53,11241,113</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,017</td>
<td>Chambers,Alabama,AL,50</td>
<td>&quot;frontier status, population size, poverty, age&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,019</td>
<td>Cherokee,Alabama,AL,64</td>
<td>&quot;frontier status, population size, poverty, age&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,021</td>
<td>Chilton,Alabama,AL,32</td>
<td>&quot;frontier status, population size, poverty, age&quot;</td>
<td>37,41744</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,023</td>
<td>Choctaw,Alabama,AL,66</td>
<td>&quot;frontier status, population size, poverty, age, populati</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,025</td>
<td>Clarke,Alabama,AL,51</td>
<td>&quot;frontier status, population size, poverty, age, populatic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01,027</td>
<td>Clay,Alabama,AL,63</td>
<td>&quot;frontier status, population size, poverty, age, populati</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fixed Record-length Format
Motivation

• Data.gov and similar initiatives provide data in CSV format or similar

• W3C Government Linked Data Working Group Charter
  – “The mission ... is to provide standards and other information which help governments around the world publish their data as effective and usable Linked Data using Semantic Web technologies.”

• Machine-actionability - intended for program use
Example at data.gov
Existing Approaches

• **CSV on the Web Working Group Charter** (W3C)

• Linkage
  – **Linked CSV** (Jeni Tennison)
  – **CSV linked data** (Quoderat)

• Description
  – Common Format and MIME Type for Comma-Separated Values (CSV) Files, [RFC 4180](https://tools.ietf.org/html/rfc4180)
  – **csv**: a vocabulary for describing CSV files (Rurik Thomas Greenall, Norwegian University at Trondheim)

• Representations
  – **URI design for RDF conversion of CSV-based data** (Tim Lebo, Gregory Todd Williams)
  – **CSV2RDF Application** (Ivan Ermilov, Sören Auer, Claus Stadler)

**Research results:**
  – Intended for different purposes
  – Description approaches not sufficient
PHDD – First Ideas
PHDD – UML Model
PHDD - Overview

General approach is not really new, just a complete set of properties for the most common cases.

Structure

- **Table** – the rectangular data file
  
  [disco::DataFile, dcat::Distribution]
  
  – **TableStructure** - common properties plus specific ones for delimited and fixed columns

- **Column** - common properties plus specific ones for delimited and fixed columns
  
  [disco::Variable]
Table Structure

- FixedRecordLength
  - recordLength : xsd:positiveInteger [0..1]

- Delimited
  - delimiter : xsd:string
  - textQualifier : xsd:string [0..1]
  - consecutiveDelimitersAsOne : xsd:boolean
  - namesOnFirstRow : xsd:boolean
  - firstDataLine : xsd:positiveInteger

StructuredBy

1

0..*

TableStructure

- characterSet : xsd:string
- defaultDecimalSeparator : xsd:string [0..1]
- defaultDigitGroupSeparator : xsd:string [0..1]
- defaultLanguage : xsd:string [0..1]
- defaultLocale : xsd:string [0..1]
- defaultDecimalPositions : xsd:positiveInteger [0..1]
- newLine : xsd:string = CRLF

0..*

0..*

0..*
Relationship to other RDF Vocabularies

class externalVocabularies

- dcat::Distribution
  - dcat::distribution
    - dcat::Dataset
      - dcat::dataset
        - dcat::Catalog

- Table
  - isStructuredBy
    - 0..*
      - TableStructure
        - column
          - 0..*
            - Column
              - 1..*

- disco::DataFile
  - disco::dataFile
    - disco::LogicalDataSet
      - disco::containsVariable
        - disco::Variable

- ´owl:equivalentClassº
  - disco::dataFile
  - disco::LogicalDataSet
    - disco::containsVariable
      - disco::Variable
  - ´owl:equivalentClassº

- ´owl:equivalentClassº
  - disco::DataFile
  - disco::LogicalDataSet
    - disco::containsVariable
      - disco::Variable
  - ´owl:equivalentClassº
Relationship to DDI XML

• Mapping to DDI XML Specifications
  – DDI Codebook 2.*
    • approx. half of the properties of PHDD
  – DDI Lifecycle 3.*
    • almost all properties of PHDD
Relationship of DDI Specifications

XML Schema
- DDI Codebook 2.*
- DDI Lifecycle 3.*

OWL/RDF
- PHDD
- Discovery
- XKOS

Future
- DDI 4 Model

- XML Schema Representation
- OWL/RDF Representation
Acknowledgements

• Contributions by
  – Larry Hoyle (Institute for Policy & Social Research, University of Kansas)
  – Richard Cyganiak (DERI - Digital Enterprise Research Institute)
Further Information

• Development repository of PHDD

• DDI Alliance RDF Vocabularies