

PHDD – An RDF Vocabulary for the Physical Data Description

Work in Progress

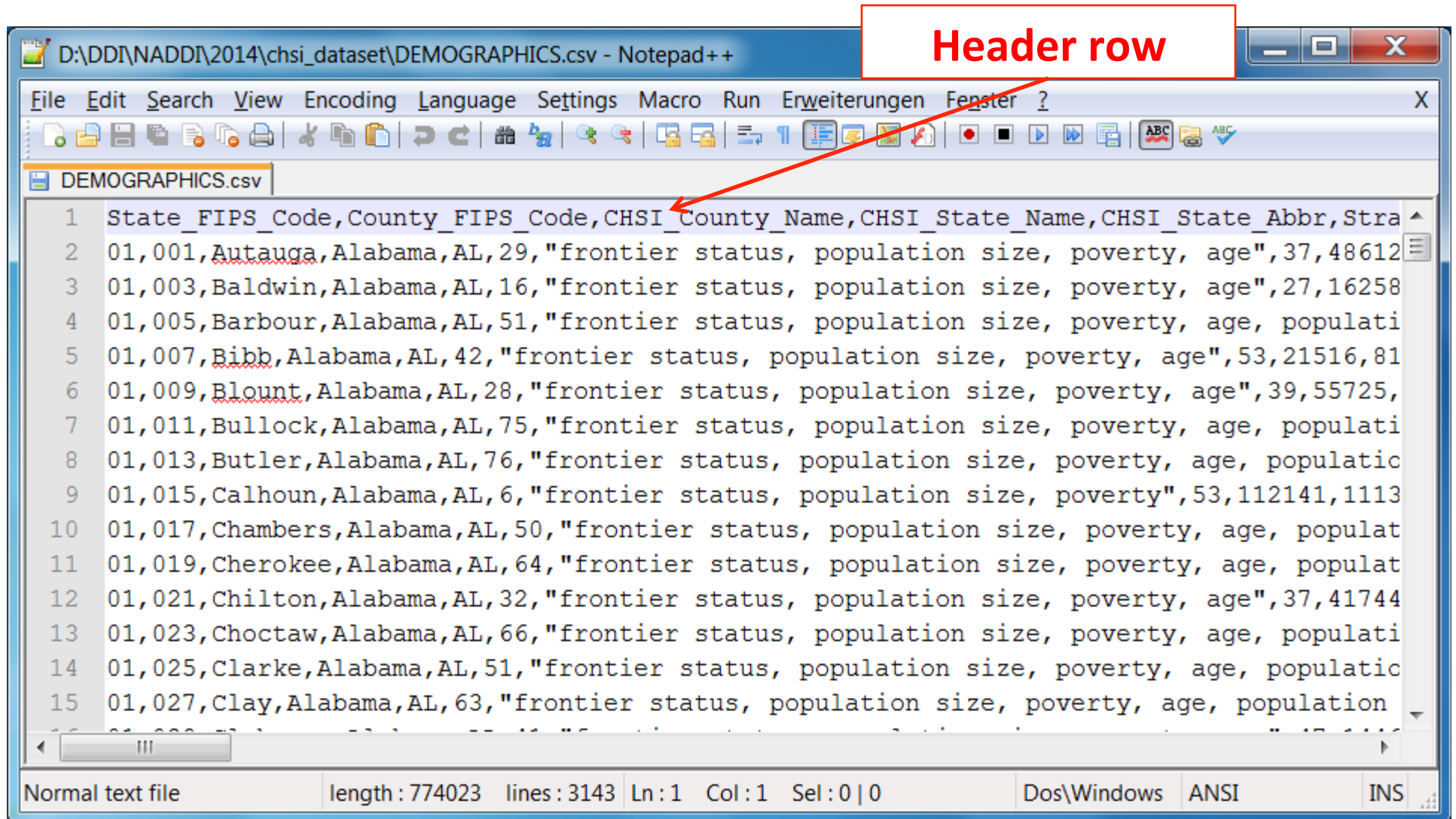
Joachim Wackerow and Thomas Bosch

(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

Character-separated Values



Header row

```
1 State_FIPS_Code,County_FIPS_Code,CHSI_County_Name,CHSI_State_Name,CHSI_State_Abbr,Stratification,Population,Poverty
2 01,001,Autauga,Alabama,AL,29,"frontier status, population size, poverty, age",37,48612
3 01,003,Baldwin,Alabama,AL,16,"frontier status, population size, poverty, age",27,16258
4 01,005,Barbour,Alabama,AL,51,"frontier status, population size, poverty, age, populati
5 01,007,Bibb,Alabama,AL,42,"frontier status, population size, poverty, age",53,21516,81
6 01,009,Blount,Alabama,AL,28,"frontier status, population size, poverty, age",39,55725,
7 01,011,Bullock,Alabama,AL,75,"frontier status, population size, poverty, age, populati
8 01,013,Butler,Alabama,AL,76,"frontier status, population size, poverty, age, populatic
9 01,015,Calhoun,Alabama,AL,6,"frontier status, population size, poverty",53,112141,1113
10 01,017,Chambers,Alabama,AL,50,"frontier status, population size, poverty, age, populat
11 01,019,Cherokee,Alabama,AL,64,"frontier status, population size, poverty, age, populat
12 01,021,Chilton,Alabama,AL,32,"frontier status, population size, poverty, age",37,41744
13 01,023,Choctaw,Alabama,AL,66,"frontier status, population size, poverty, age, populati
14 01,025,Clarke,Alabama,AL,51,"frontier status, population size, poverty, age, populatic
15 01,027,Clay,Alabama,AL,63,"frontier status, population size, poverty, age, population
```

Normal text file length : 774023 lines : 3143 Ln : 1 Col : 1 Sel : 0 | 0 Dos\Windows ANSI INS

Fixed Record-length Format

Flat File Sample Wizard: Record Types for Fixed Length Files


Indicate the start and end position of the field to scan for record types.

Start position: End position:

	Type Value	Record Name
<input type="checkbox"/>	E	RECORD1
<input type="checkbox"/>	P	RECORD2

File: d:\BIRD\SampleData\MEmpascii.dat

Field lengths:



```
E003715415309061987014000000IRENE HIRSH 108500206600308800412500
P00371501152000011620000010100050000000070000015000200001330007500055000660007700
P003715021520000216200000102000300000000080000012000180001200006500044000750005500
P003715031520000316200000103000500000000090000013000170001100005500033000650006600
```

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

The screenshot shows a Mozilla Firefox browser window with the following elements:

- Browser Title Bar:** Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer - Data.gov - Mozilla Firefox
- Menu Bar:** Datei, Bearbeiten, Ansicht, Chronik, Lesezeichen, Extras, Hilfe
- Address Bar:** catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obe
- Breadcrumbs:** / Organizations / U.S. Department of Health & ... / Community Health Status ...
- Left Sidebar:**
 - HealthData.gov Federal** logo
 - U.S. Department of Health & Human Services**
 - Share on Social Sites:** Google+, Twitter, Facebook
- Main Content Area:**
 - Dataset** tab
 - Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer**
 - Description:** Community Health Status Indicators (CHSI) to combat obesity, heart disease, and cancer are major components of the Community Health Data Initiative. This dataset provides key health indicators for local communities and encourages dialogue about actions that can be taken to improve community health (e.g., obesity, heart disease, cancer). The CHSI report and dataset was designed not only for public health professionals but also for members of the community who are interested in the health of their community. The CHSI report contains over 200 measures for each of the 3,141 United States counties. Although CHSI presents indicators like deaths due to heart disease and cancer, it is imperative to understand that behavioral factors such as obesity, tobacco use, diet, physical activity, alcohol and drug use, sexual behavior and others substantially contribute to these deaths.
 - Data and Resources** section with a **CSV** icon and text: **CSV** chsi_dataset.zip
 - Download** button

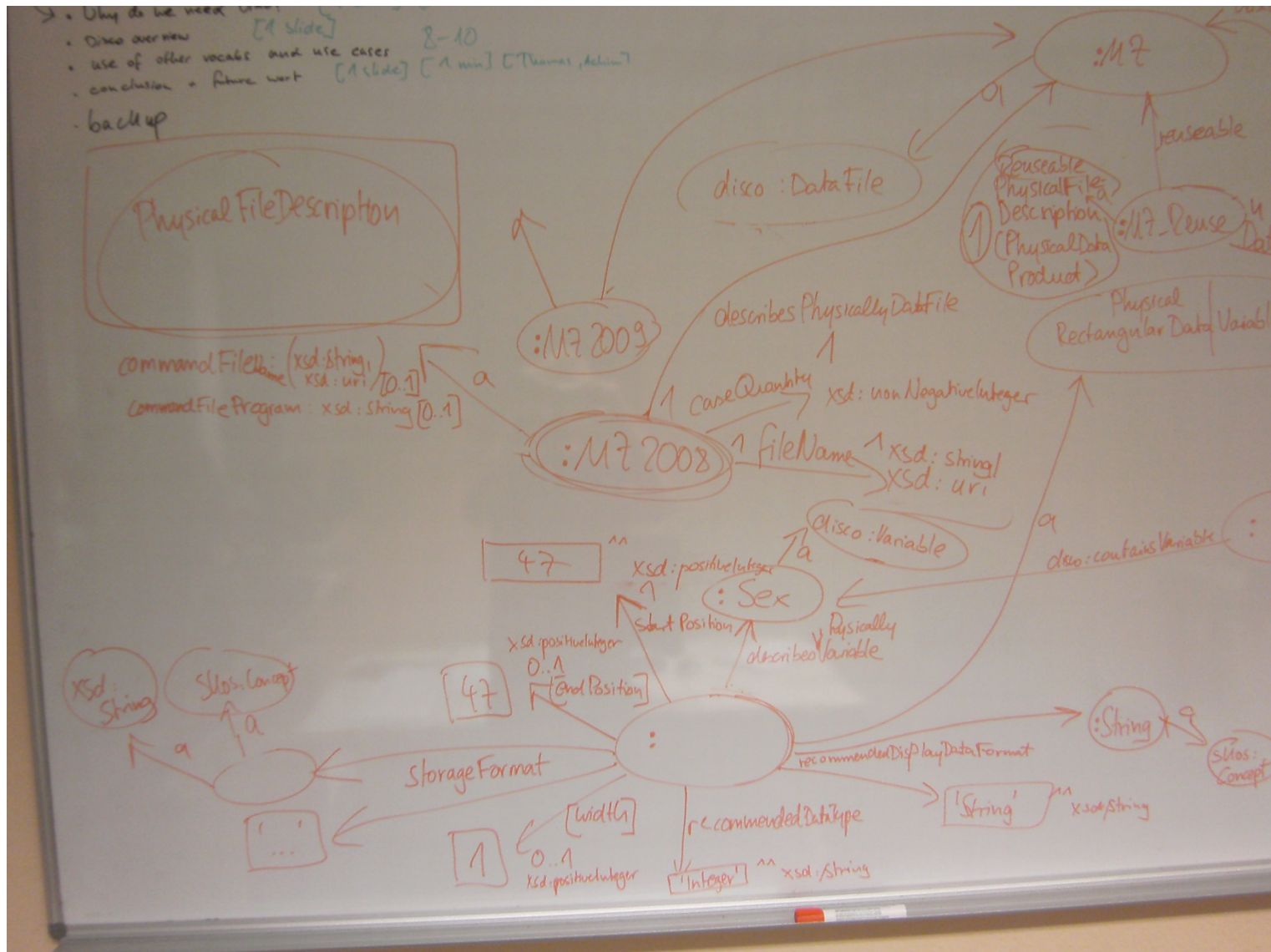
Existing Approaches

- [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](#)
 - [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)

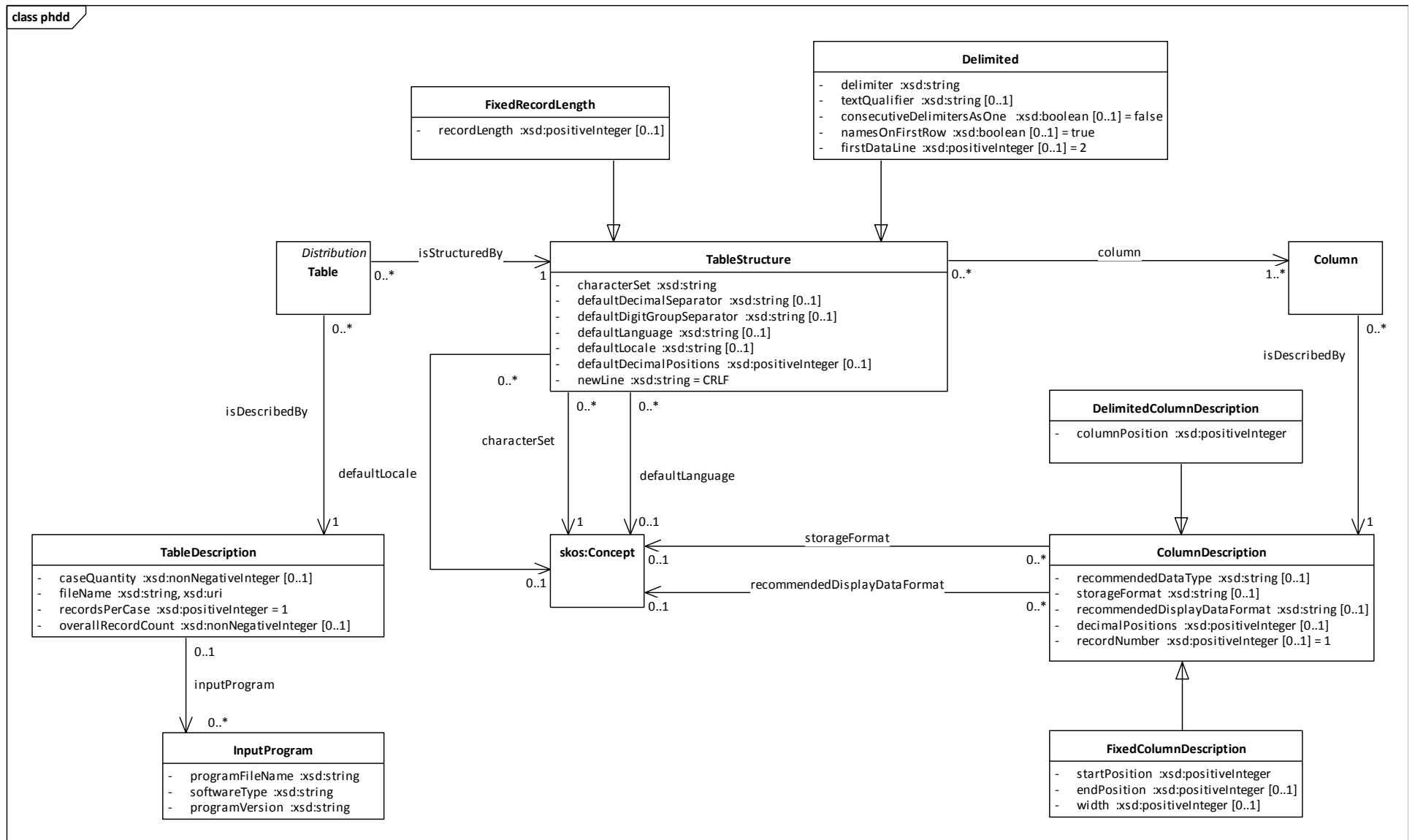
Research results:

- 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

Delimited

FixedRecordLength

recordLength :xsd:positiveInteger [0..1]

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

StructuredBy

TableStructure

1

0..*

- 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..*

Column Description

DelimitedColumnDescription

- columnPosition :xsd:positiveInteger

ColumnDescription

- recommendedDataType :xsd:string [0..1]
- storageFormat :xsd:string [0..1]
- recommendedDisplayDataFormat :xsd:string [0..1]
- decimalPositions :xsd:positiveInteger [0..1]
- recordNumber :xsd:positiveInteger [0..1] = 1

FixedColumnDescription

- startPosition :xsd:positiveInteger
- endPosition :xsd:positiveInteger [0..1]
- width :xsd:positiveInteger [0..1]

storageFormat

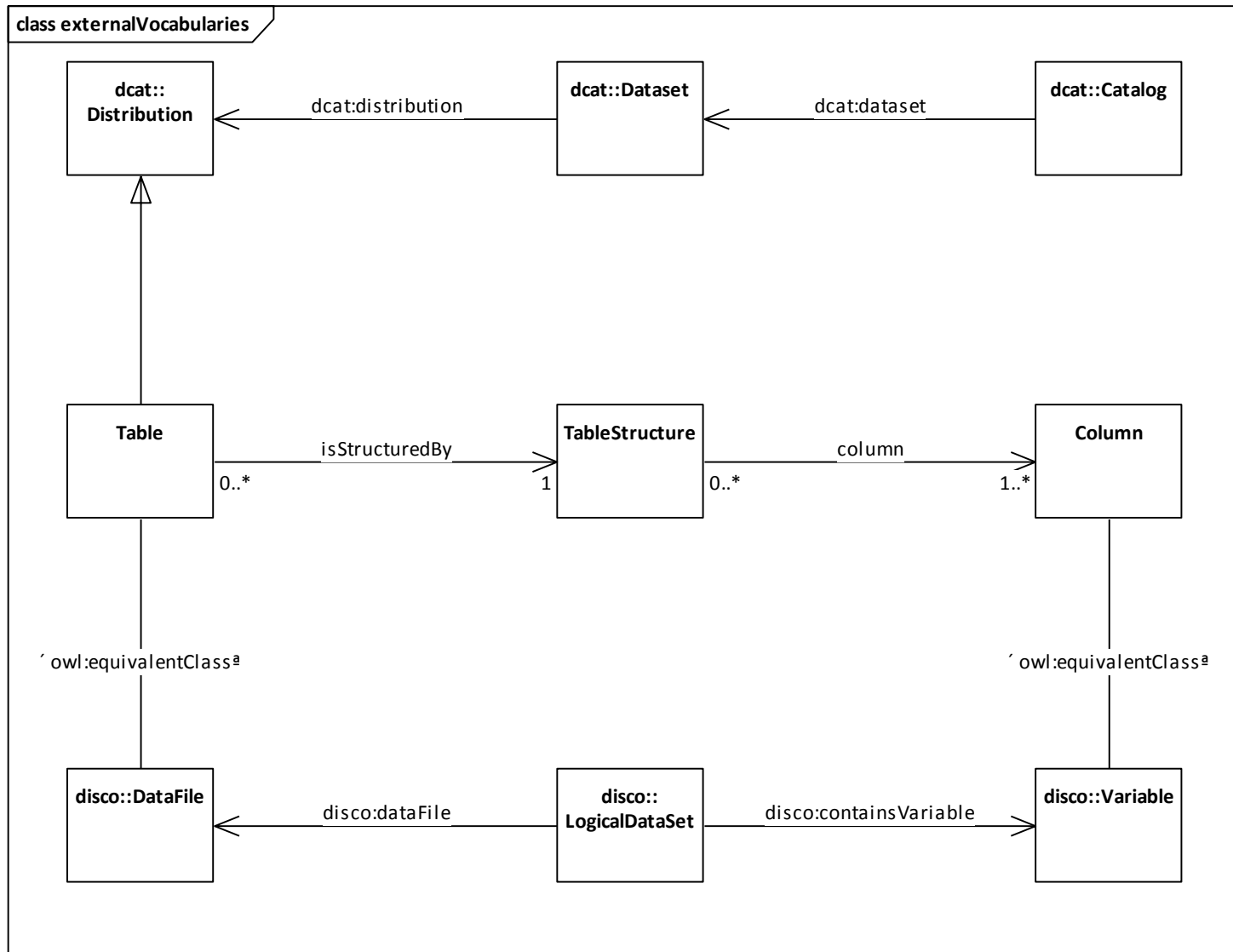
0..*

recommendedDisplayDataFormat

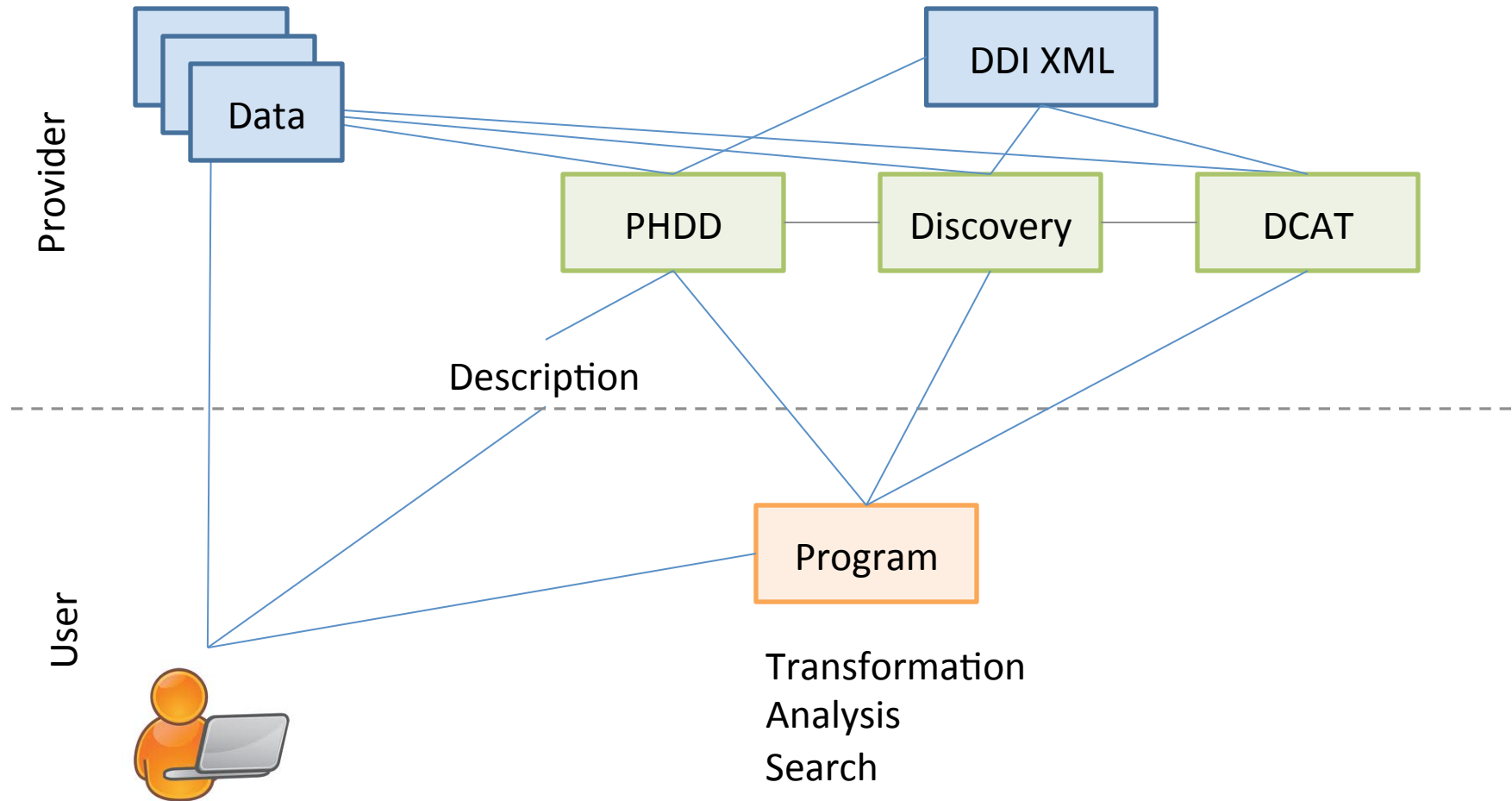
0..*

1

Relationship to other RDF Vocabularies



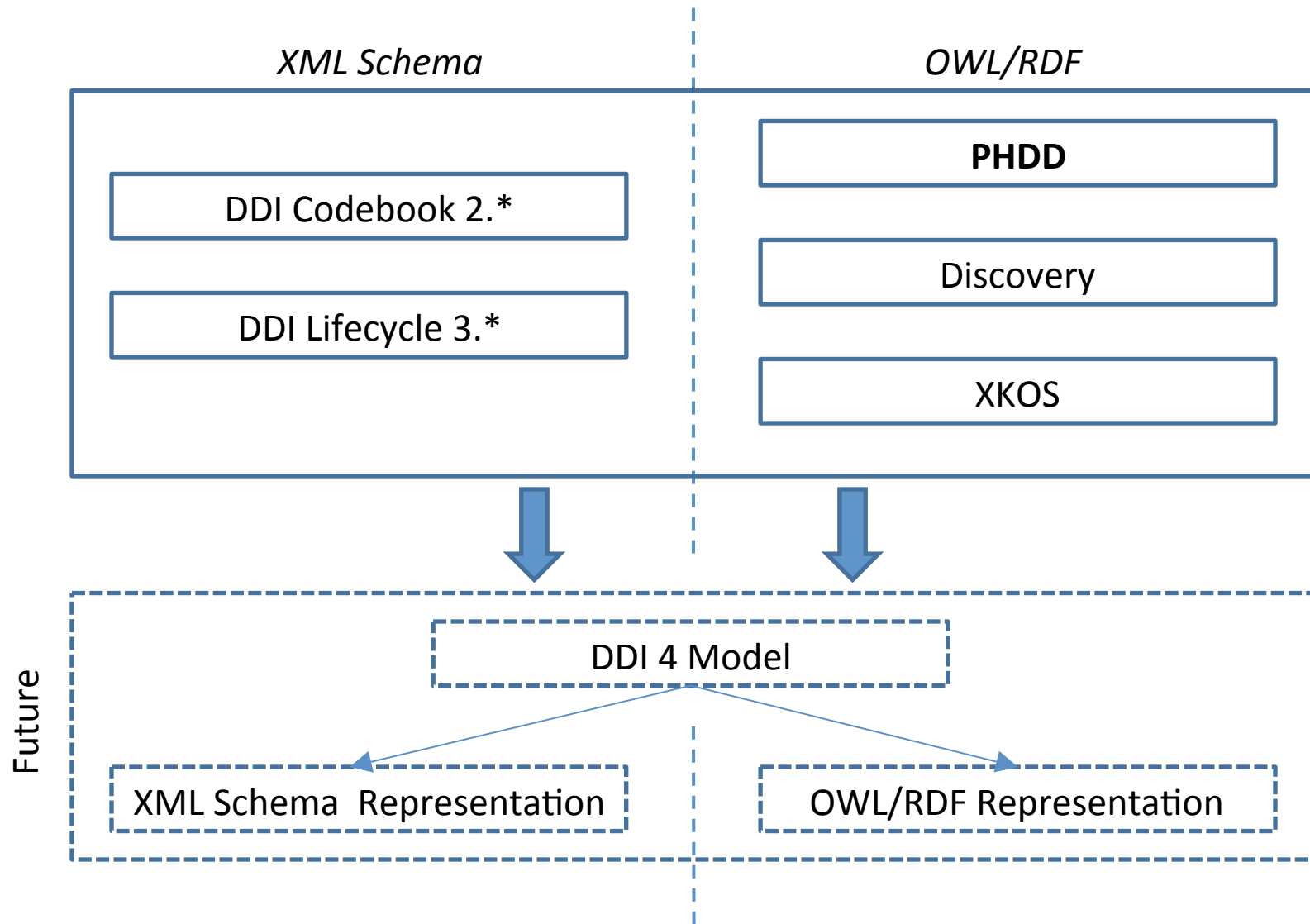
Usage Scenarios



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



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
 - <https://github.com/linked-statistics/physical-data-description>
- DDI Alliance RDF Vocabularies
 - <http://www.ddialliance.org/Specification/RDF>