PHDD – An RDF Vocabulary for the Physical Data Description

Work in Progress

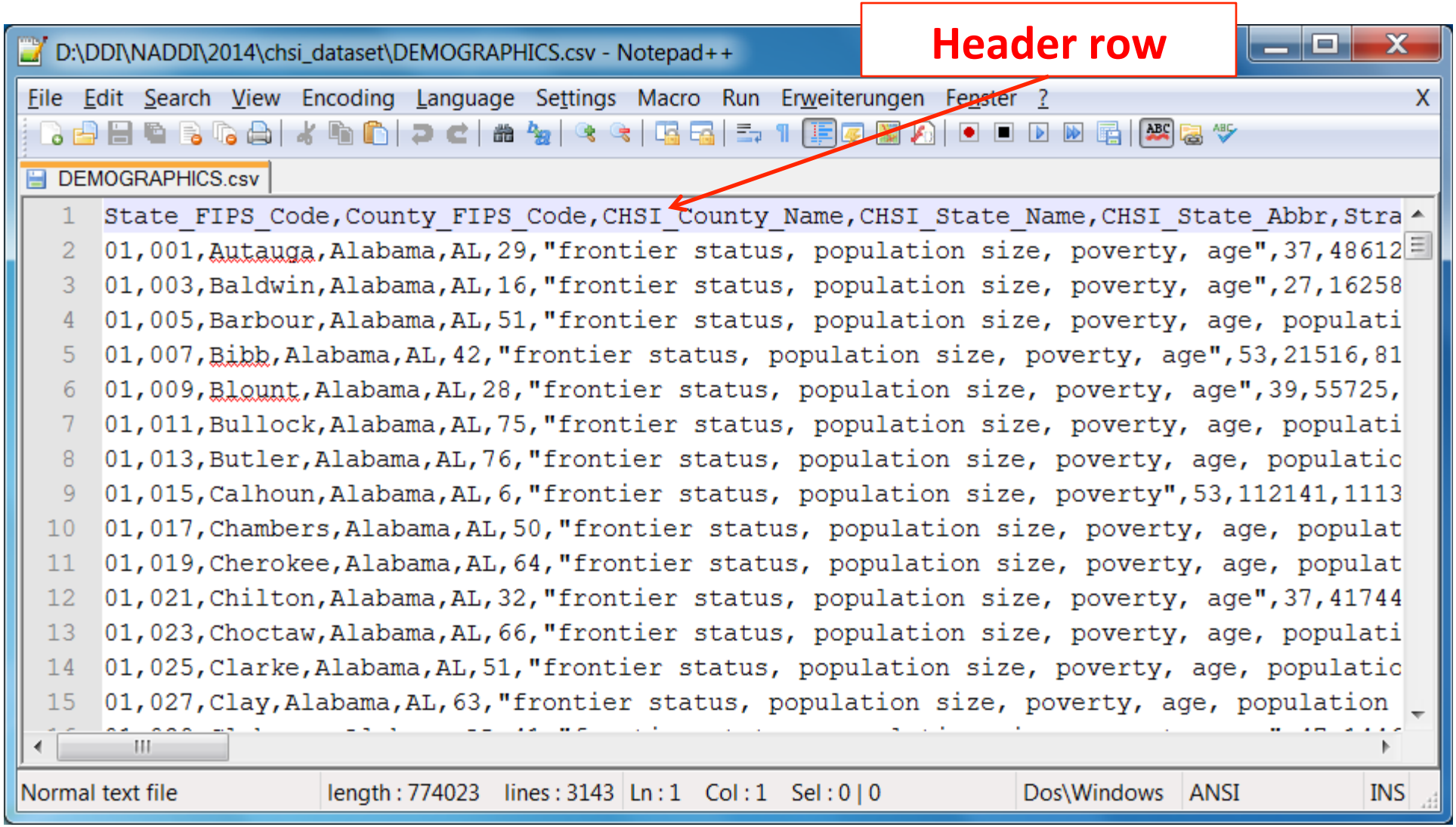
Joachim Wackerow and Thomas Bosch

(both GESIS – Leibniz Institute for the Social Sciences)

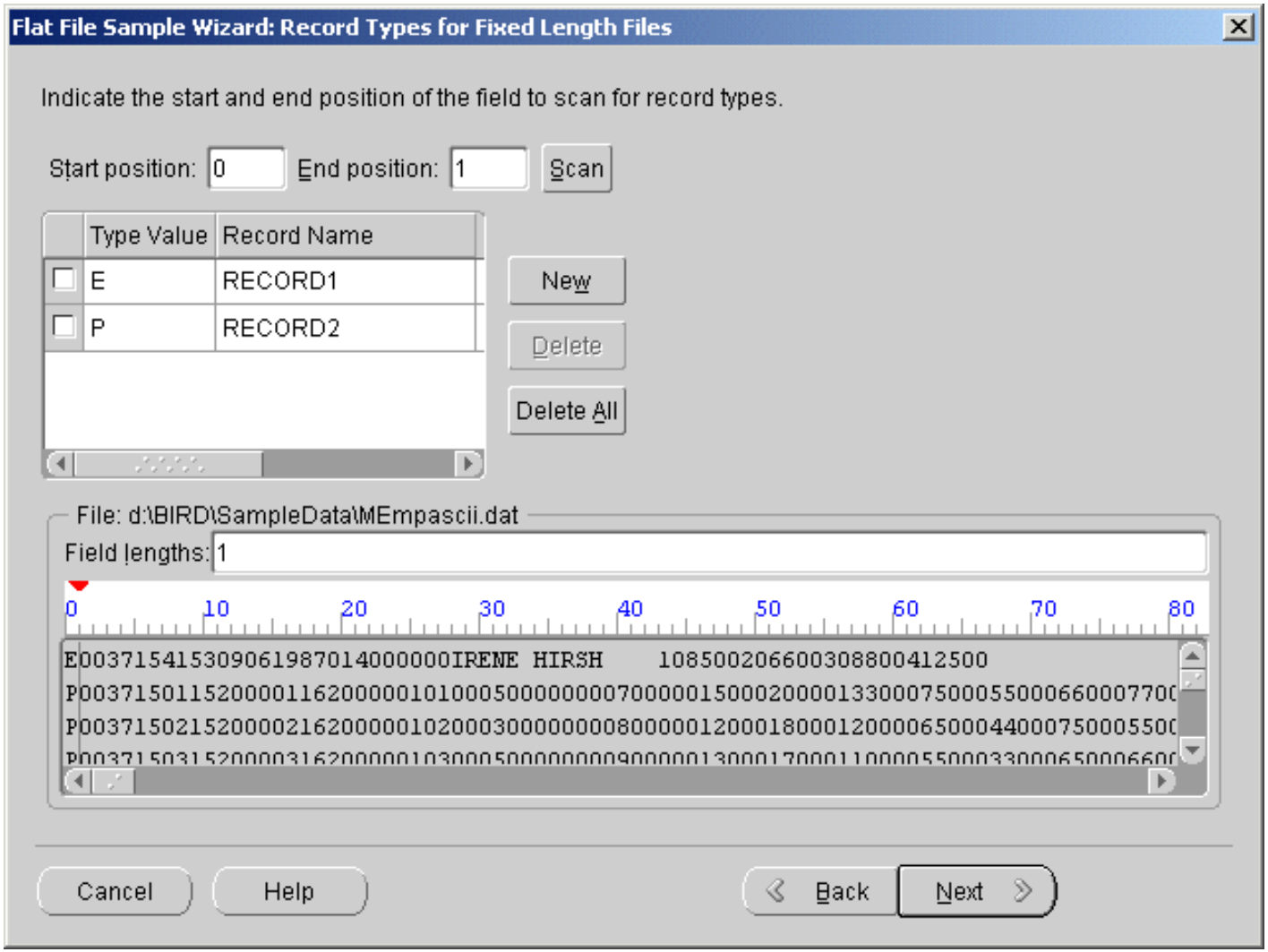
What is it?

- 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



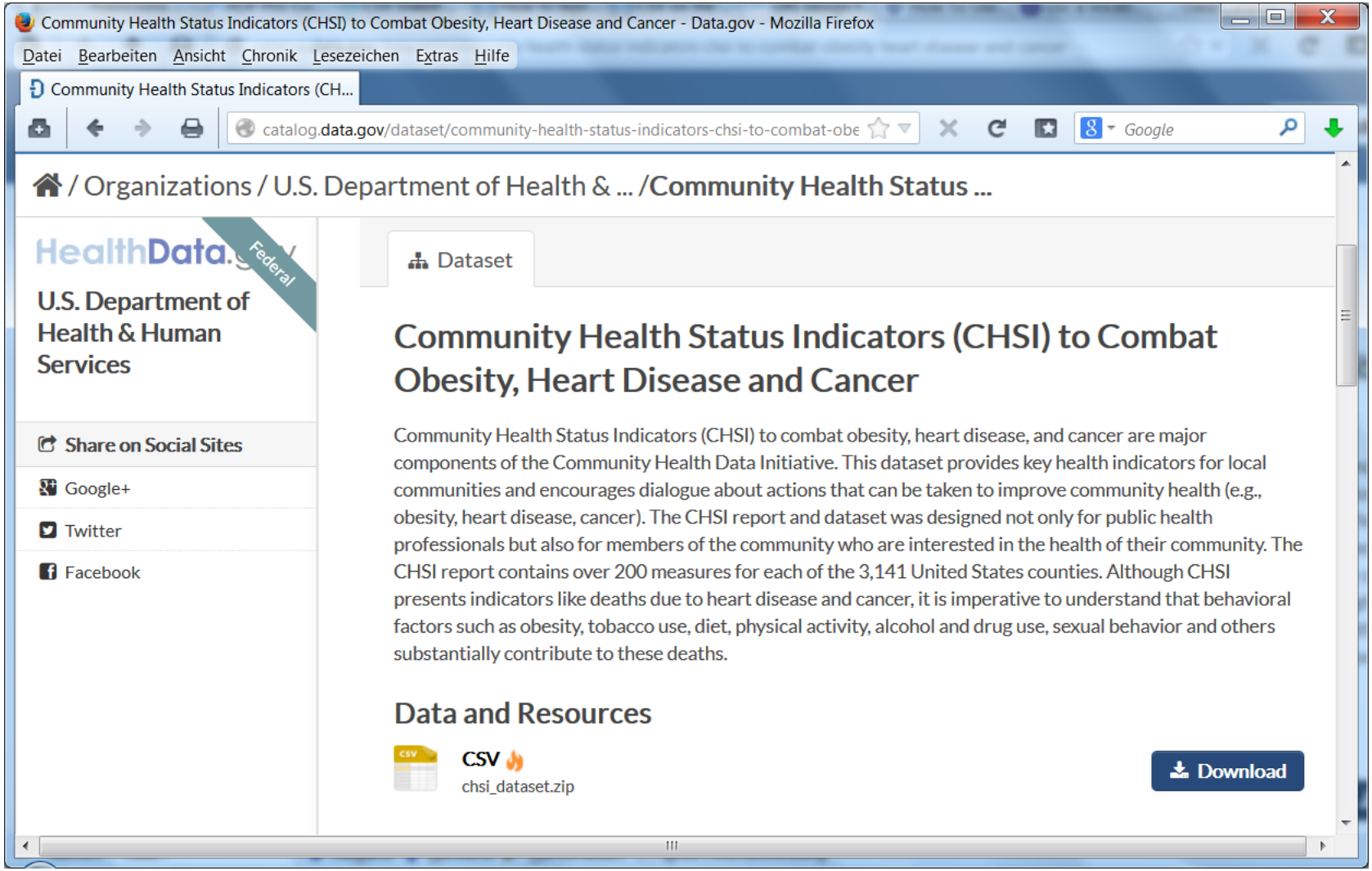
Fixed Record-length Format



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



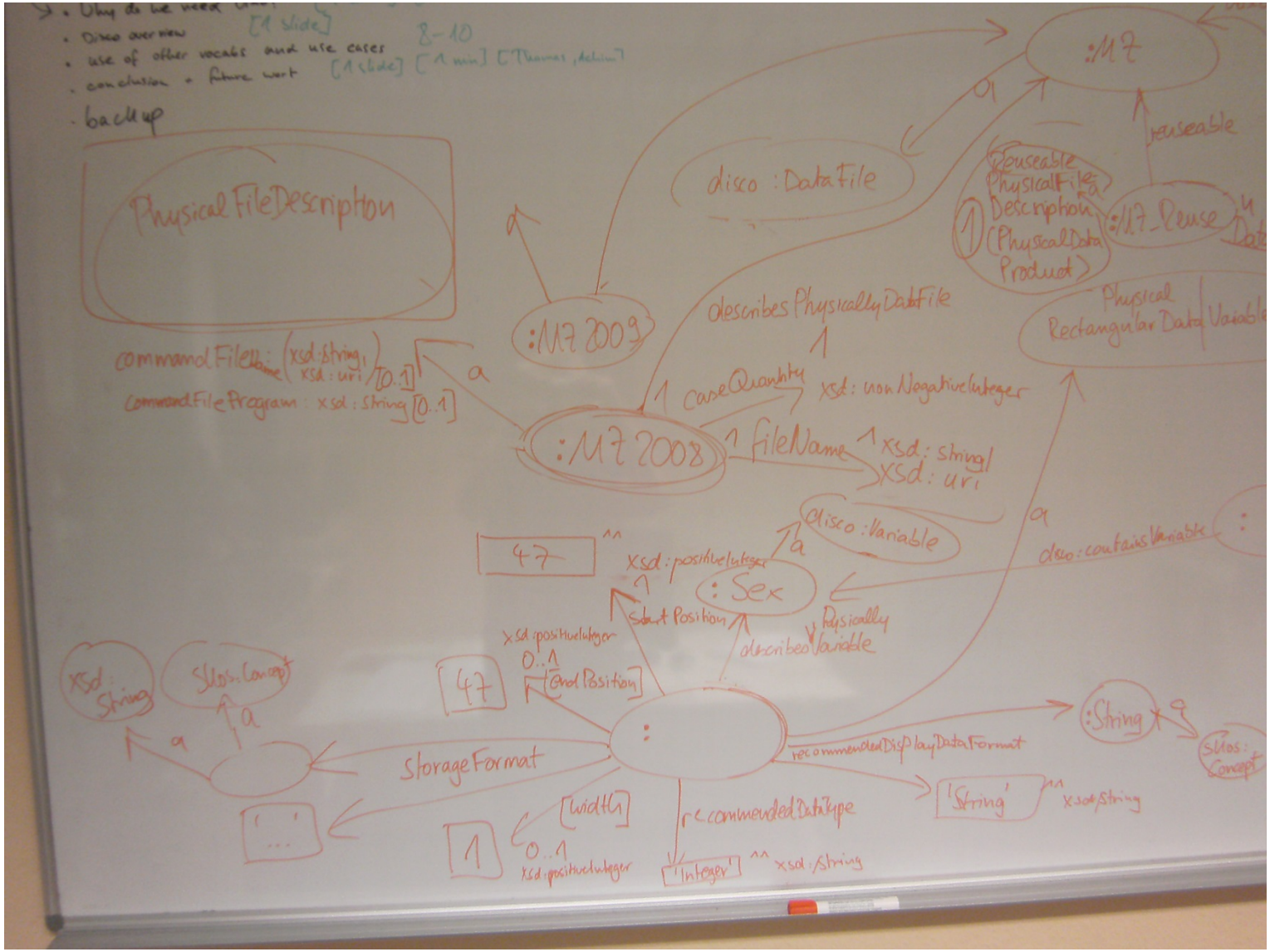
Existing Approaches

- 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
  - 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)

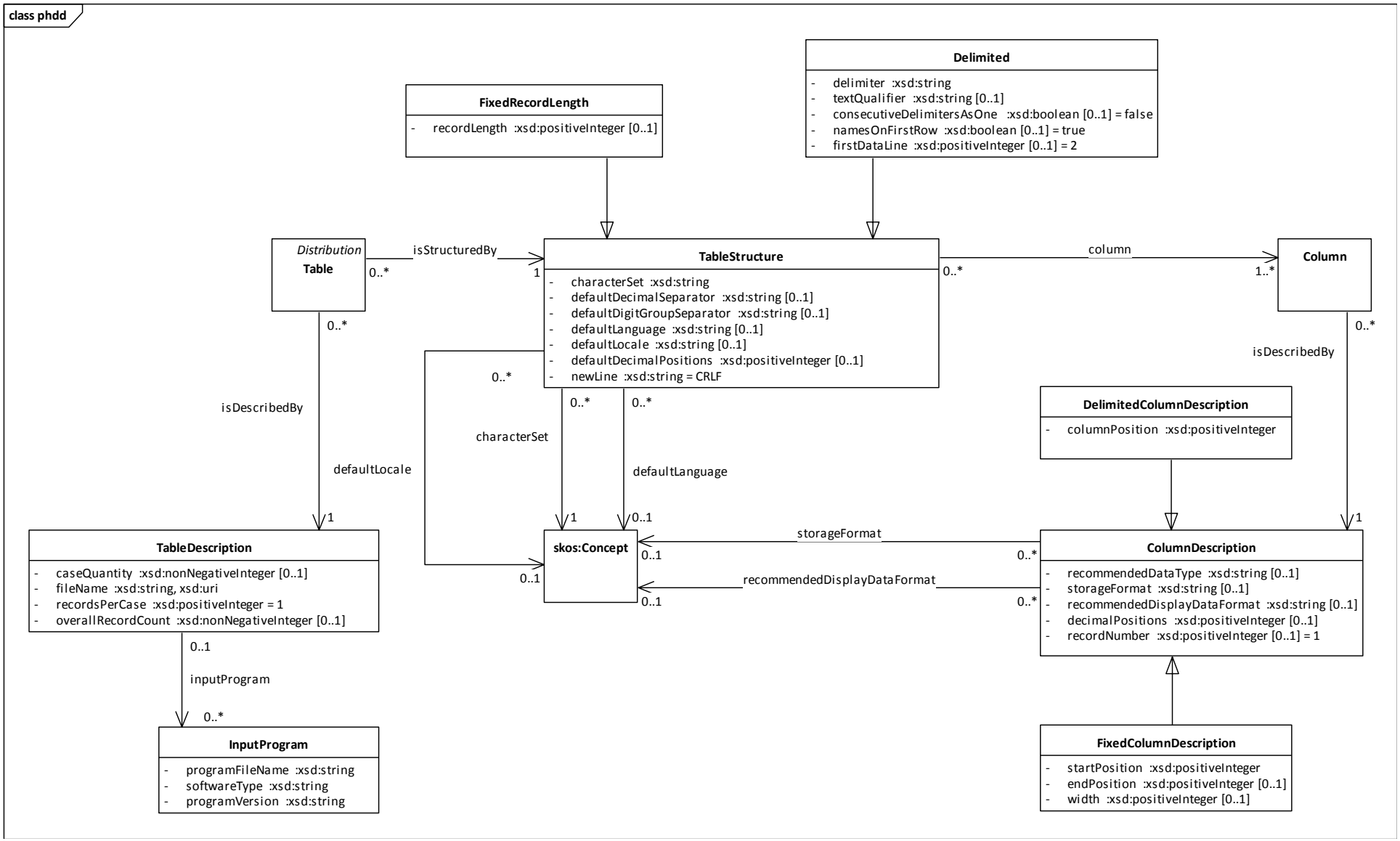
Research results:

- 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

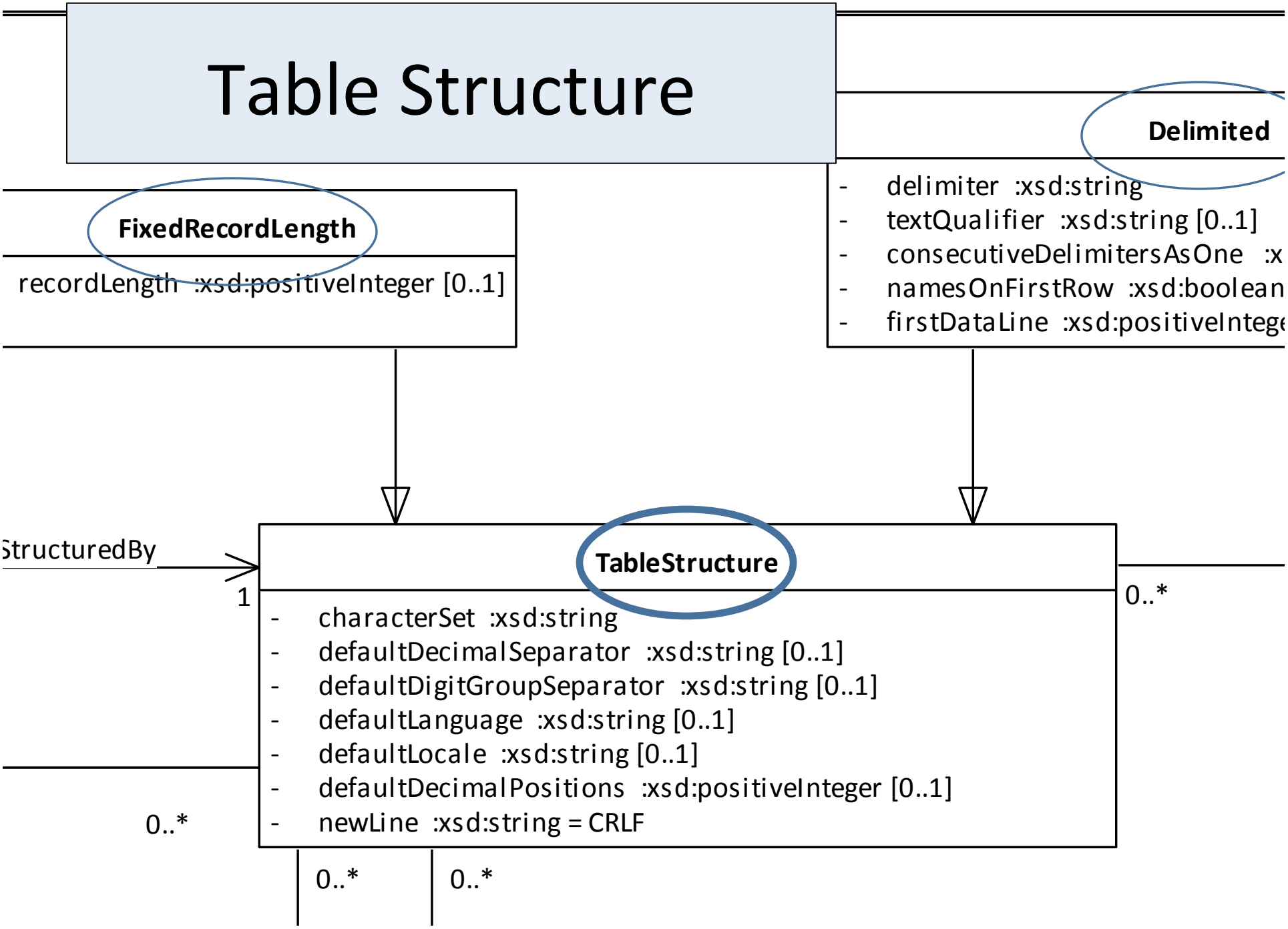
- **Table** – the rectangular data file

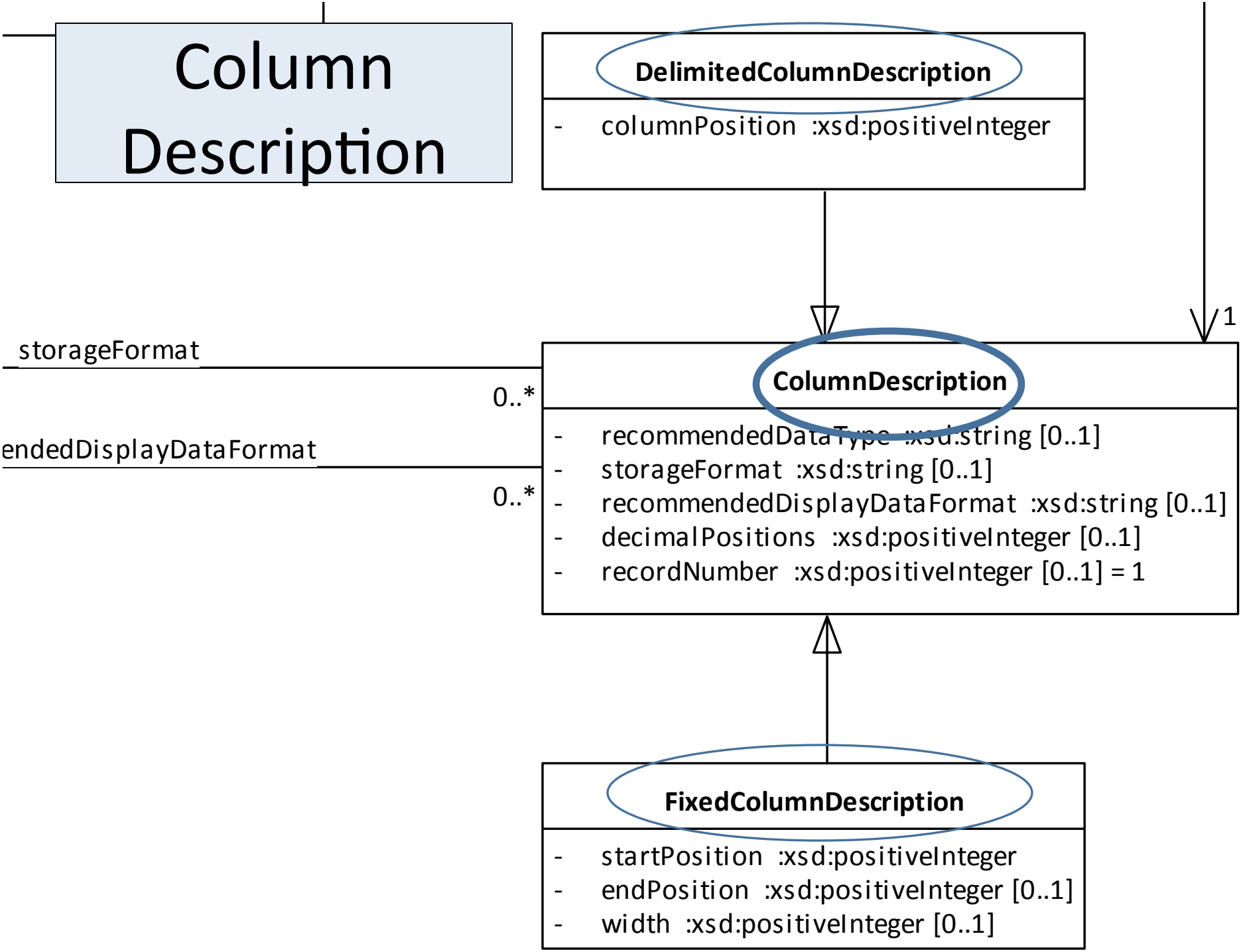
  [disco::DataFile, dcat::Distribution]

  - **TableStructure** - common properties plus specific ones for delimited and fixed columns
    - **Column** - common properties plus specific ones for delimited and fixed columns

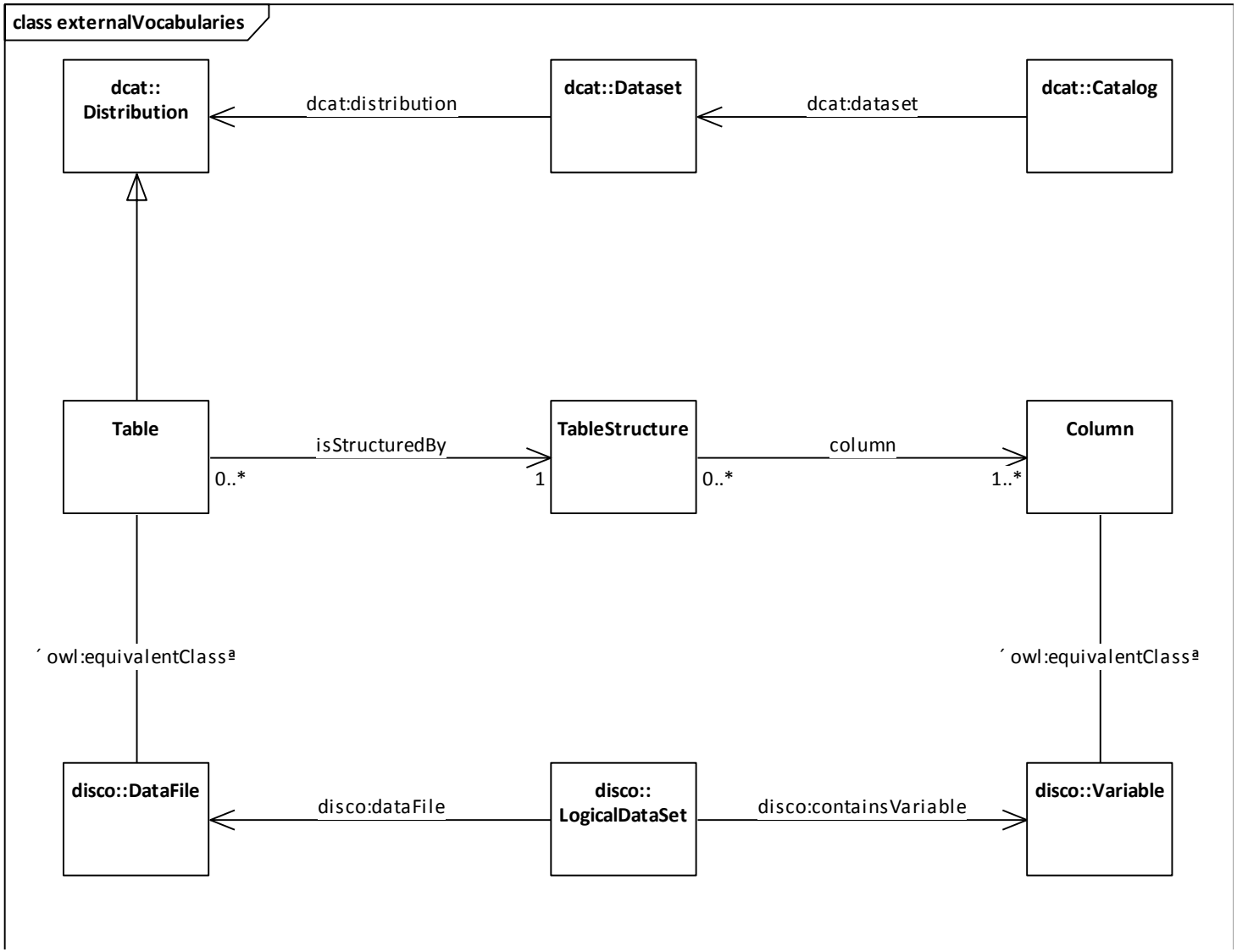
      [disco::Variable]

Table Structure

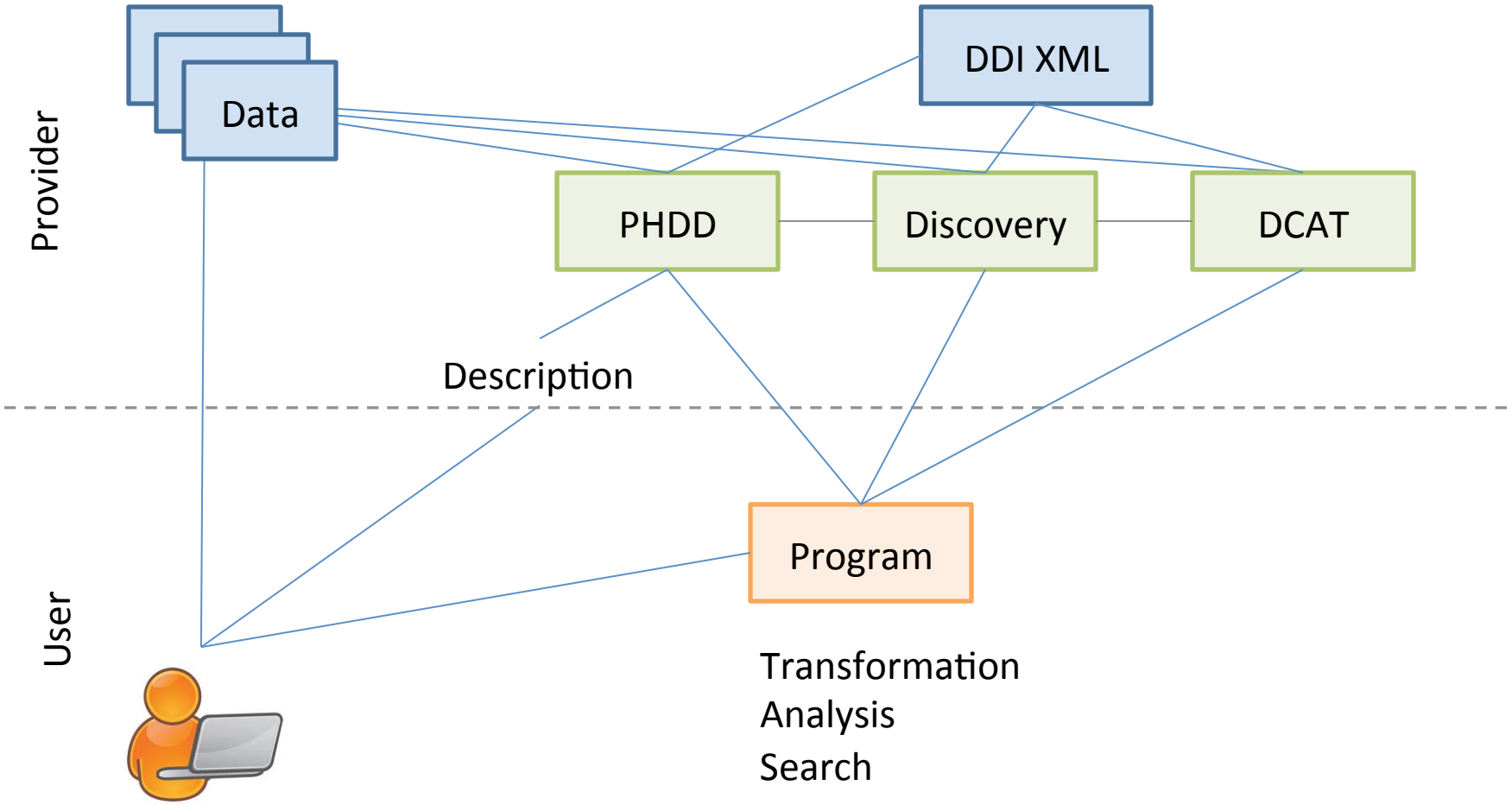



Column Description
DelimitedColumnDescription
- columnPosition :xsd:positiveInteger
1
storageFormat
0..*
endedDisplayDataFormat
0..*
ColumnDescription
- recommendedDataType :xsd:string [0..1]
- storageFormat :xsd:string [0..1]
- recommendedDisplayDataFormat :xsd:string [0..1]
- decimalPositions :xsd:positiveInteger [0..1]
- recordNumber :xsd:positiveInteger [0..1] = 1
FixedColumnDescription
- startPosition :xsd:positiveInteger
- endPosition :xsd:positiveInteger [0..1]
- width :xsd:positiveInteger [0..1]

Relationship to other RDF Vocabularies



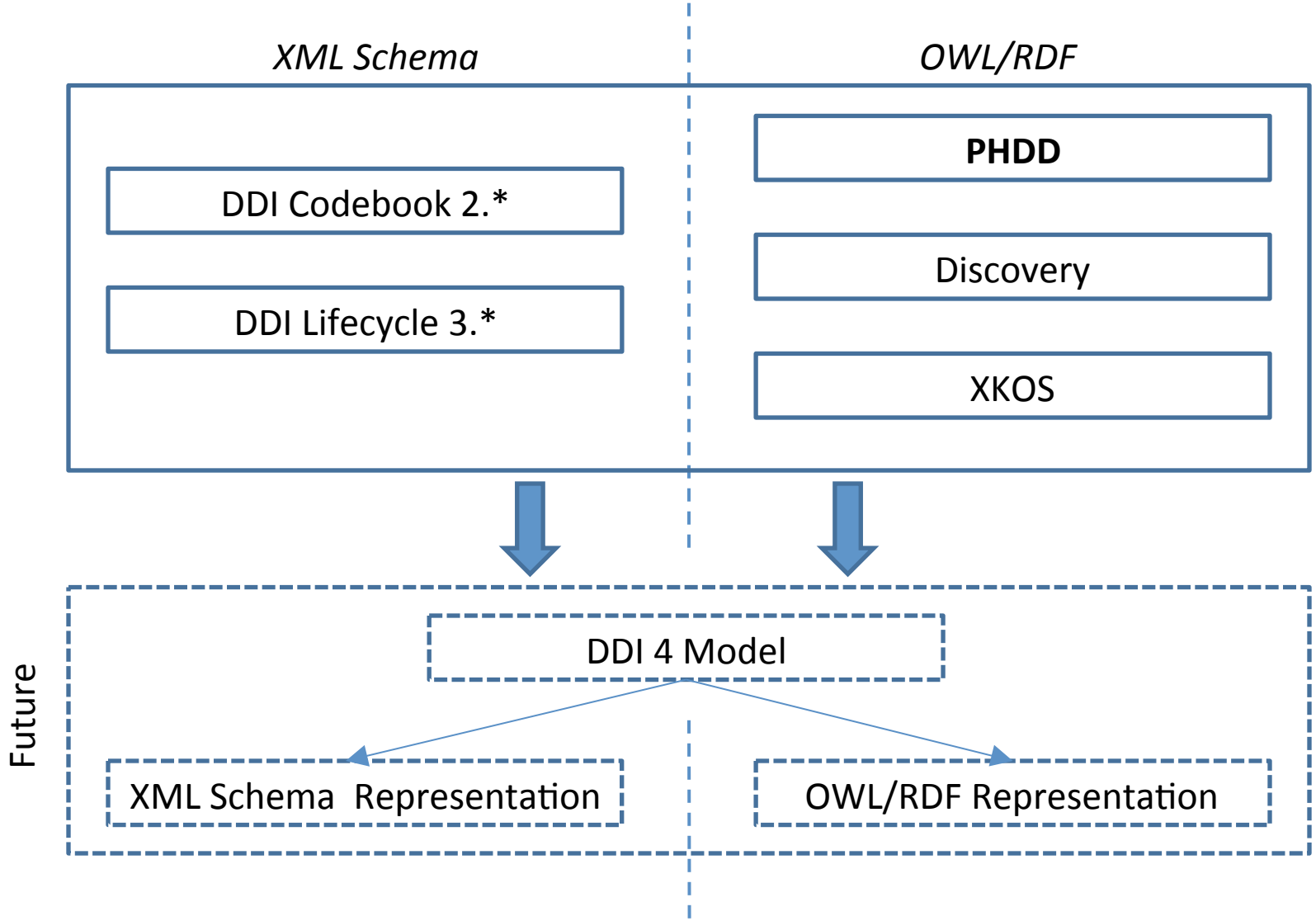
Usage Scenarios



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