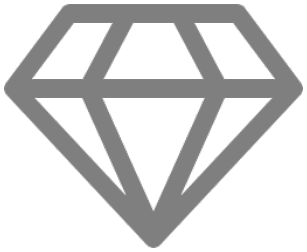


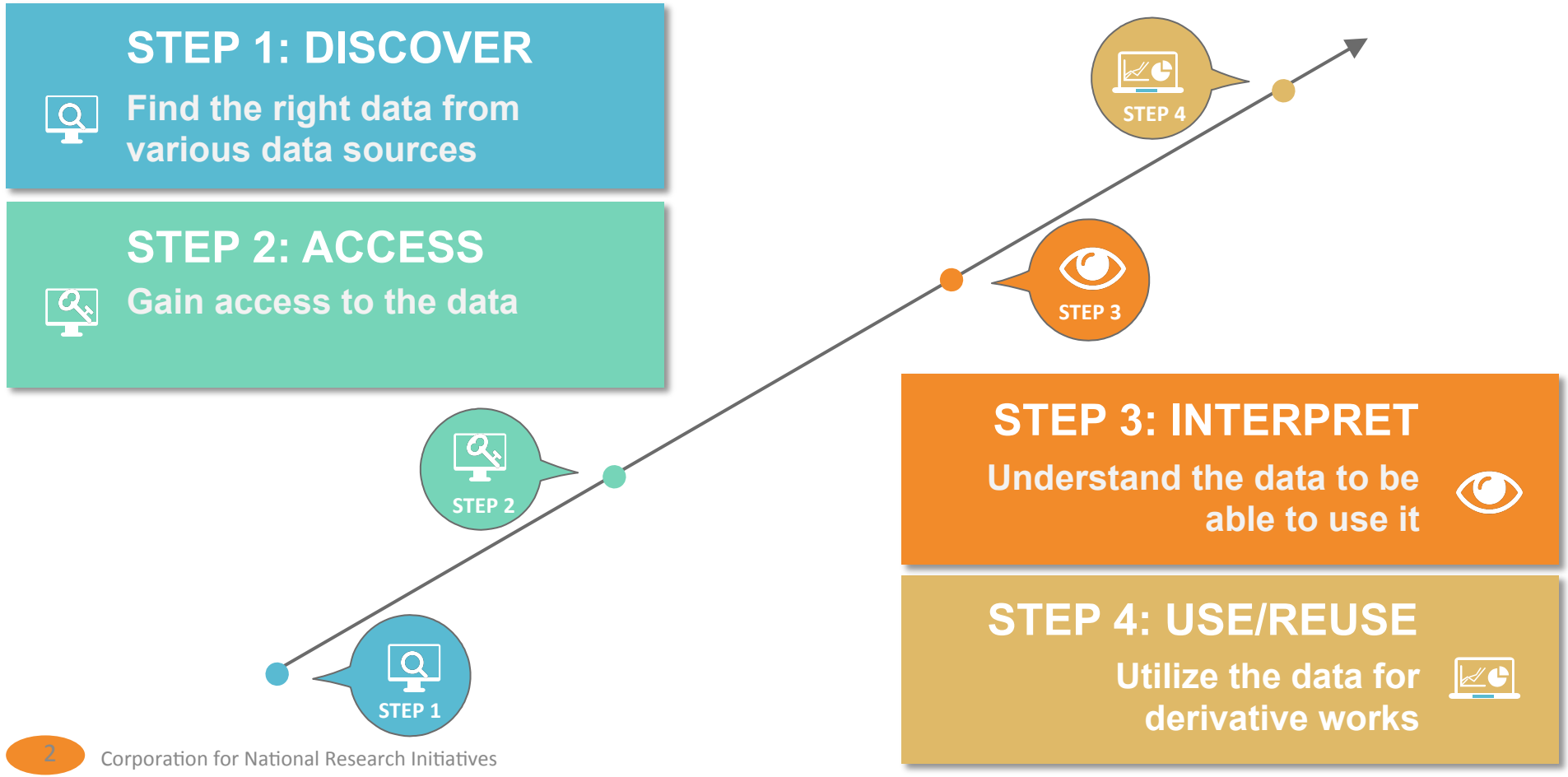
Data Typing Efforts

January 28, 2016

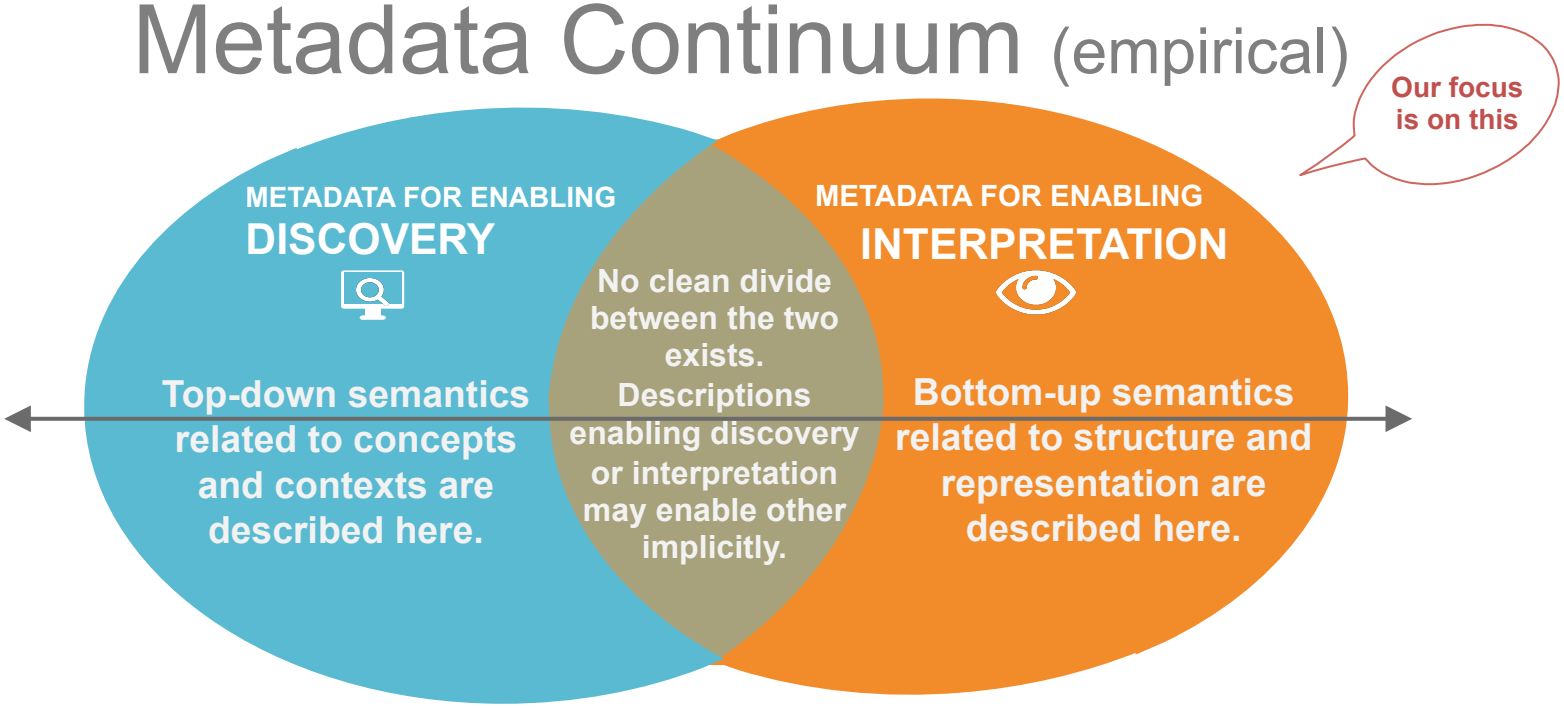
http://enrich.cordra.org/

Corporation for National Research Initiatives

(a typical) Journey of a data consumer



Metadata Continuum (empirical)



Discovery metadata is used to answer queries of the form:
I am looking for "2015 US housing affordability data".

Interpretation metadata is used to answer queries of the form:
Where are "Area Median Income" values in the dataset?
Are those under the column "AMI"?
If so, how is AMI calculated?

Problem space

The “interpretability” problem we are trying to solve is

How do you describe data in a way that will help consumers understand and begin to use/reuse the data.

Problem space

- Solving the interpretability problem (for even humans) in an interoperable way across all domains is extremely hard.
- Challenges exist mainly because of conceptual and contextual differences between users, especially those from varying domains.

Problem space

- Conceptual differences are primarily vocabulary related; select reasons for vocabulary issues include:
  - *Vocabulary silos*: some words have no meaning outside of a domain.
  - *Homonyms*: same word is used, but meaning varies by domain.
  - *Synonyms*: same meaning is intended, but expressed differently in different domains.
- Contextual issues arise because critical information lies outside of data.
  - e.g., the economic data is only about minorities - not the entire population - but the dataset itself does not carry that information.

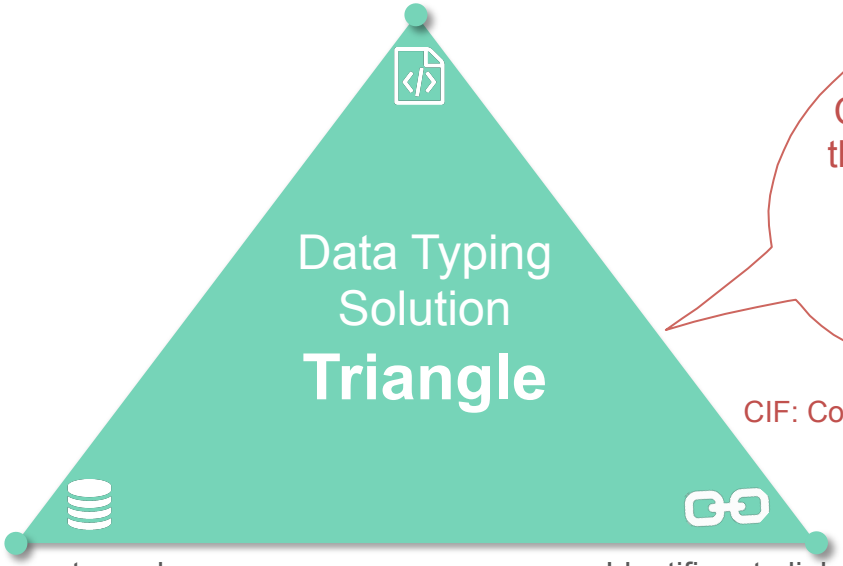
Solution space

- Identifying a single solution that works for all domains and all consumers is hard.
  - Attempting to produce specific solutions targeting specific domains and specific consumer profiles is more tractable.
- However, certain components would be common across domain-specific solutions as well as different implementations.

Solution space

Any solution to this problem would include certain common pieces:

A powerful descriptive language to describe datasets
e.g., ISO 11179, RDF/OWL, etc.
We call the descriptions “data type records”.



CIF includes these three pieces; CIF is being researched and developed currently.

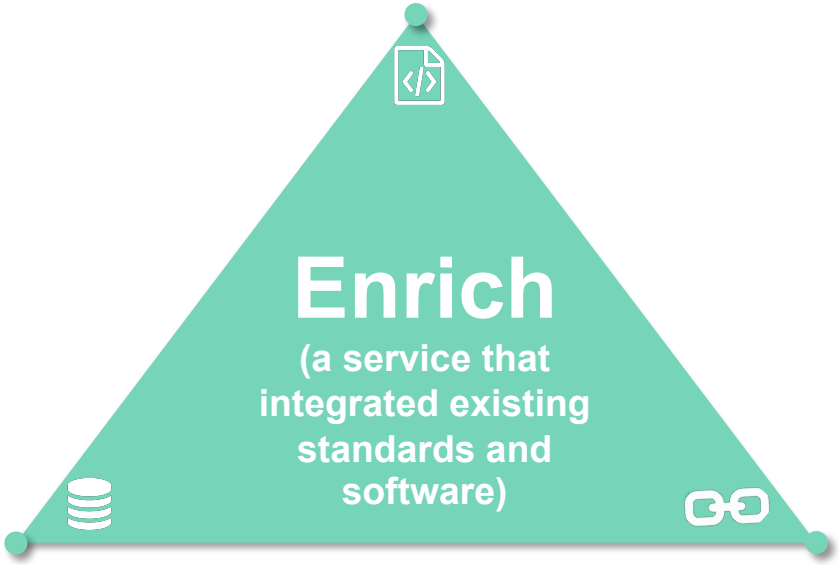
CIF: Commerce Interoperability Framework

A widely known system where such descriptions are held;
we call such system a “data type registry”

Identifiers to link datasets and data type records,
e.g., Handles, DOIs, PURLs, etc.

One approach: Enrich

Data Type records are expressed in a custom structure using JSON.
Next step is to express using ISO 11179.



Cordra.org software is used for storing auto-generated data type records, harvested metadata about datasets, and concepts & units harvested from NLM.

Handles from handle.net are used to link datasets, data type records, and concepts & units together.

Enrich

- **Enrich** is a short experimental effort that produced a few desirable results using existing standards and software:
  - An easy-to-install software integrated with standard components listed in the *Solution Triangle*.
  - A *Concept and Measurement Unit Registry (DTR)* of entries covered in ISO 2955, ANSI X3.50, ISO+ (HL7 extensions) (credit: NLM).
  - A process that auto-generates drafts of data type records for tabular datasets.
  - A capability that uses the generated data type records to produce richer datasets.
  - API and UI support.

Enrich: data typing

- Auto-generation of data type records is based on machine learning algorithms and parsing logic.
- An example auto-generated type record looks like this:
  - We found “15 columns” with names X, Y, etc., in the table.
  - X looks like an integer, Y like a date, etc.
  - X values range between 1 and 120, Y is formatted as YYYY-MM-DD, etc.
- Those data type records can be further edited or updated, especially with semantic details (concepts and measurement units) from the DTR, e.g.,
  - X is a temperature expressed in Fahrenheit, Z is US zip-code, etc.

Enrich: constructing datasets

- The generated data type record could guide consumers to enable
  - transformation of values from, say, one format or measurement unit to another format or unit
  - joining/merging values between datasets based on commonality found between those datasets.

Enrich: value proposition

The single most value-added benefit of Enrich developed in a couple of months is that

Consumers can list all columns and their syntactic nature of each of the datasets without having to download those datasets.

Enrich: demo

- Still work in progress. A preliminary demo is available now.
  - Clean tabular datasets from data.gov are used for the demo.
  - Select datasets are manually updated to associate semantic details with the dataset.
- Prototype is available at http://enrich.cordra.org/