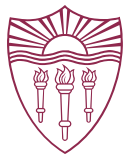
USC Viterbi
Information Sciences Institute



WDPlus: Leveraging Wikidata to Link and Extend Tabular Data

Daniel Garijo, Pedro Szekely

Information Sciences Institute and
Department of Computer Science

@dgarijov
dgarijo@isi.edu

Abundance of data sources in the Web

Raw data

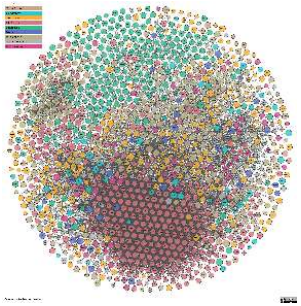
Users of data face three challenges

- How do I find relevant datasets for my problem?
- How do I augment my dataset with existing information?
- How can I share my integrated results with the community?

Popular initiatives for addressing these challenges

Google Dataset Search Beta

ckan



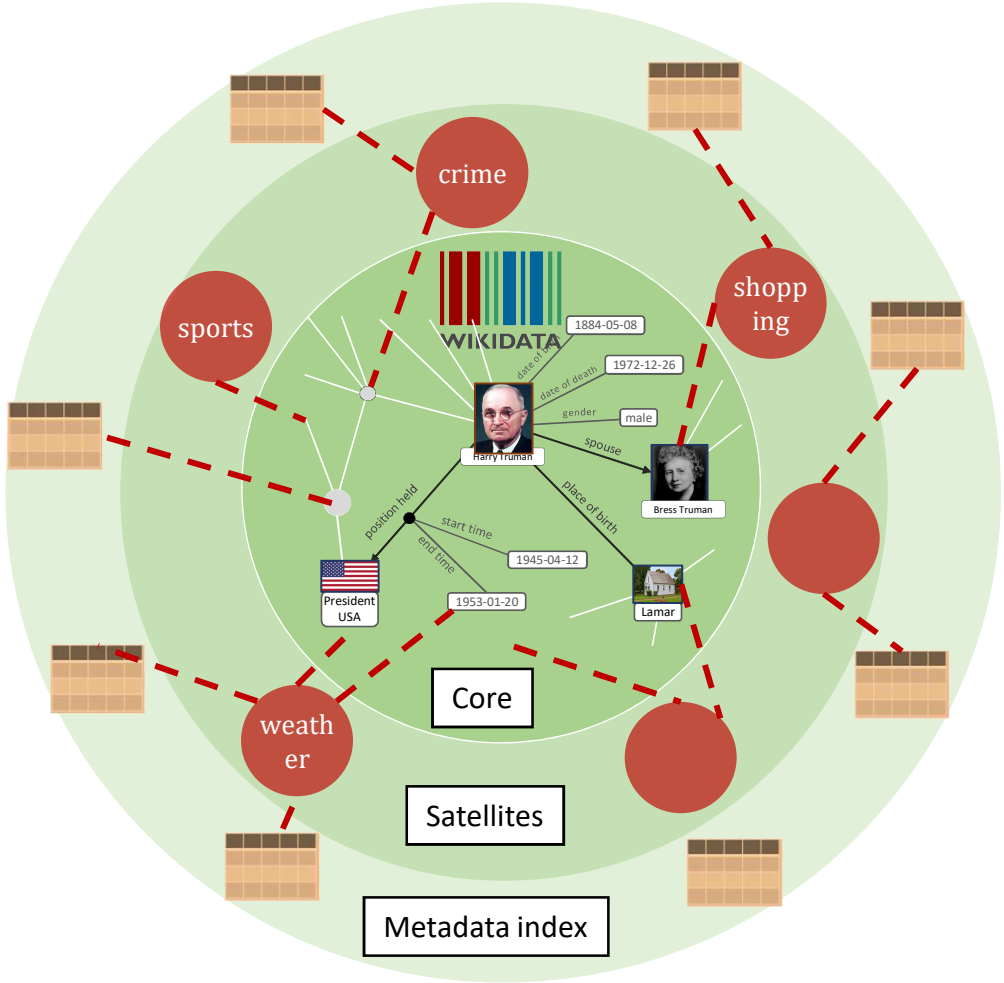
- Search individual items
  - Search is manual, based on user input
- LOD cloud of connected datasets
  - Knowledge engineers are needed to map and augment content
- ETL Frameworks (e.g, Karma, Open Refine)
  - Pipelines are custom, expertise required
  - Often not shared

Sources: https://lod-cloud.net/versions/2019-03-29/lod-cloud.png; https://panoply.io/data-warehouse-guide/3-ways-to-build-an-etl-process/

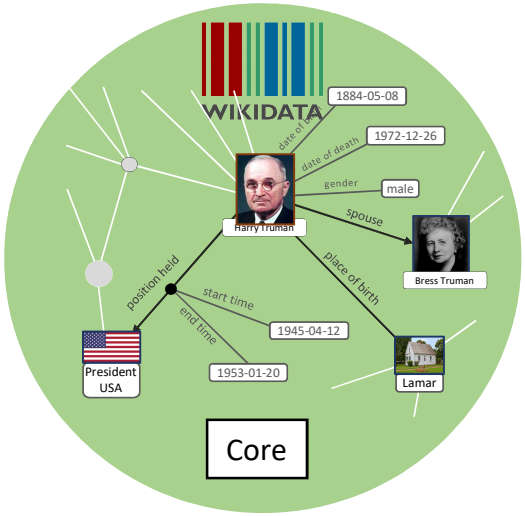
WDPlus

A framework designed to:

- Discover data on the Web
- Improve raw data to make it useful
- Search, querying dataset structure
- Download fresh data
- Combine existing dataset
- Share improved data and methods



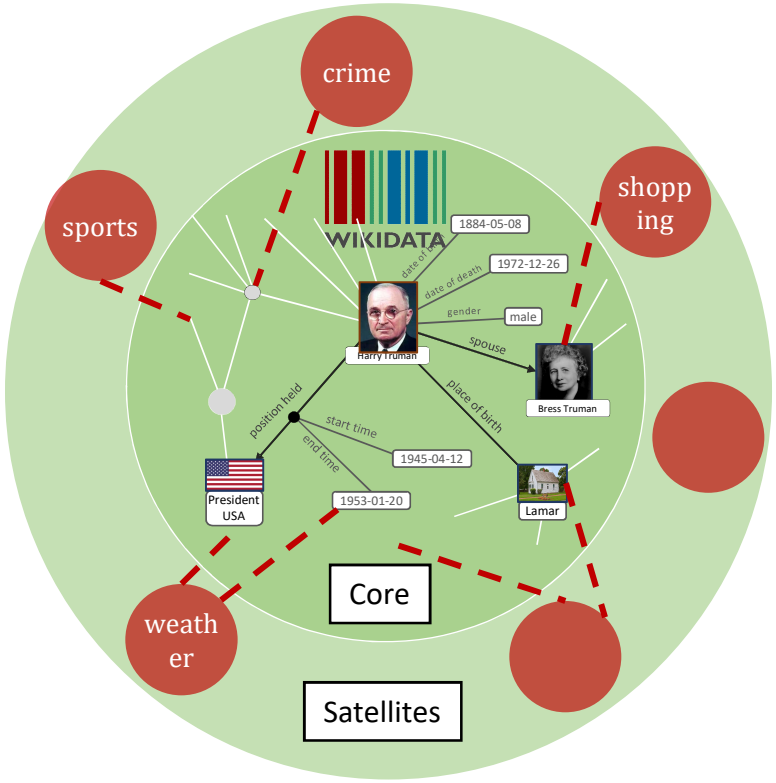
WDPlus architecture



Wikidata as a core KG

- 60 Million items
- 700 Million statements
- 20,000 + contributors
- +1 billion edits
- Collaborative!

WDPlus architecture



Satellite organization

- Detailed information on a domain
  - Crime records, sport events, etc.
- Linked to the Wikidata core
  - Link first strategy
- Custom properties and Qnodes
  - Extensions to core model
- Synchronized with core
- Decentralized
  - 1 satellite may be maintained by 1 community

WDPlus architecture

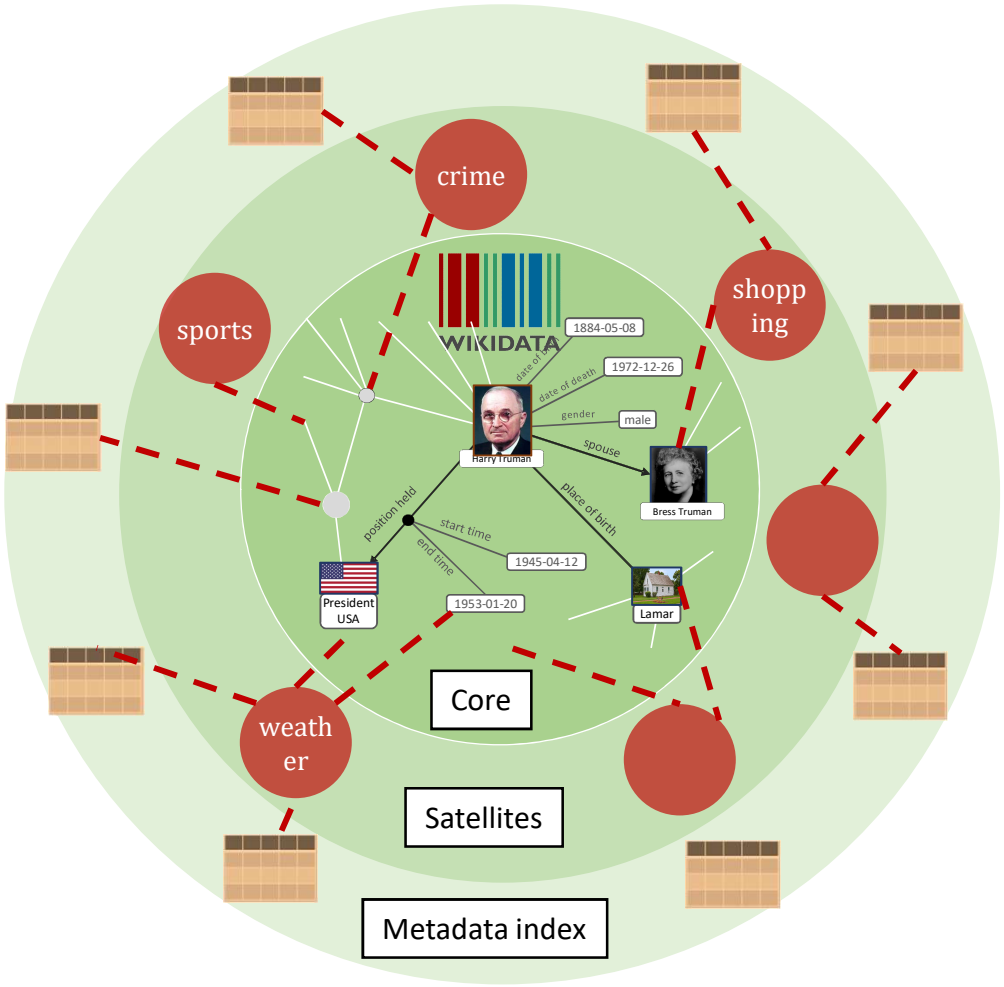
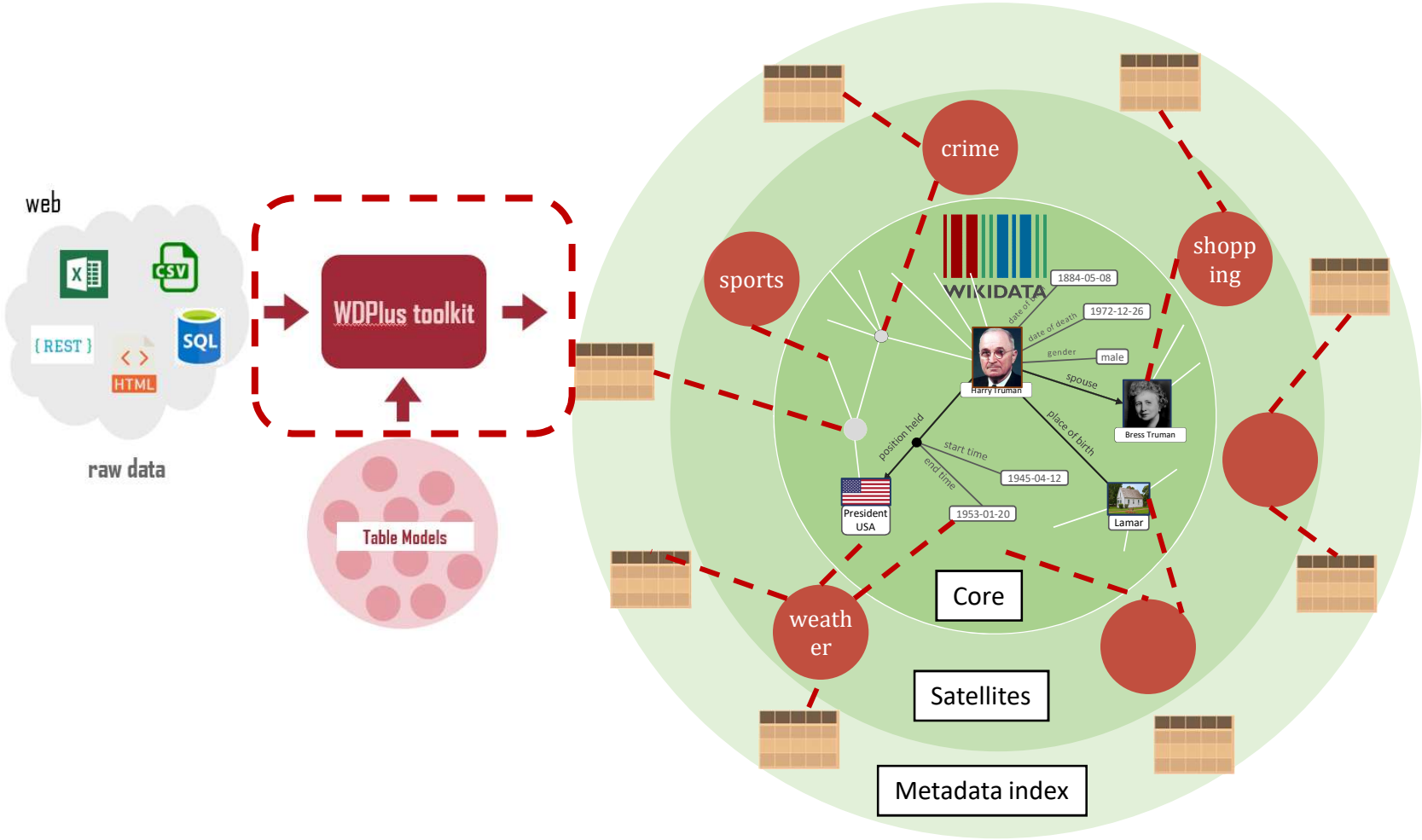


Table models

- Tables are not materialized
  - Able to become a satellite under demand
- Described in machine-readable metadata index
  - Indexing columns names and relevant instances for fast retrieval
- Link to table model is preserved

Towards WDPLus



WDPlus framework: Metadata index and table Augmentation

Search API

POST	/search	Search
POST	/search_without_data	Search by keywords or variables
POST	/wikifier	Do wikifier before search

Download API

POST	/download	Download
GET	/download/{id}	Download the dataset with given id
GET	/download_metadata/{id}	Download the dataset metadata with given id

Augment API

POST	/augment	Augment dataset

Upload API

POST	/check_upload_status	an api used to check the uploading status
POST	/upload	Upload dataset with metadata
POST	/upload/generateWD+Metadata	Generate wikified dataset and its metadata
POST	/upload/test	Test upload dataset with metadata
POST	/upload/uploadWD+Metadata	upload data and metadata

Embeddings API

GET	/embeddings/fb/{qnode}	Fetch the FB embeddings for QNODE(s)

- Search
  - Keywords, variables or content
  - Wikifier may be used in search
- Download
  - Download a dataset or its metadata
- Augment
  - Merge your dataset with contents from other datasets automatically
- Upload
  - Add new datasets (automated metadata profiling and provenance)
- Enrich
  - Header enrichment for search efficiency

WDPlus framework: T2WML

T2WML GUI v1.3 (beta) Settings

homicide_report_total_and_... Open

	A	B	C	D	E	F
1	Indicator ...	Intention...				
2						
3						
4	UNODC ...	Source		2000	2001	2002
5	Burundi	SDG	Females	1	2	2
6	Burundi	SDG	Males	4	5	6
7	Burundi	SDG	Total	5	7	8
8	Comoros	GHD Esti...	Females	2	1	n/a
9	Comoros	GHD Esti...	Males	4	n/a	8
10	Comoros	GHD Esti...	Total	6	1	8
11	Djibouti	GHD Esti...	Females	2	1	3
12	Djibouti	GHD Esti...	Males	1	1	n/a
13	Djibouti	GHD Esti...	Total	3	2	3
14	This file h...					
15	Highlight...					

{ $col: F, $row: 6 } Pinned

table-1a | table-1b | table-2a | table-2b | table-3a | table-3b

Table overview

Wikifier Open

Entity Linking

YAML Editor Open

```
statementMapping:
  region:
    - left: C
      top: 4
      bottom: value(B:/$row) = ""
      skip_row:
        - value(C/$row) = "Total"
      skip_cell:
        - value($col/$row) = "n/a"
  template:
    item: item(A/$row)
    property: P100024 # murder
    value: value($col/$row)
    qualifier:
      - property: P585 # point in time
        value: value($col/4)
        calendar: Q1985727 # Gregorian calendar
        precision: year
        time_zone: 0
      - property: P6001 # applies to people
        value: item(C/$row)
      - property: P248 # stated in
        value: item(B/$row)
```

Apply

Cell-based mapping. This mapping is saved in WDPlus for future reference

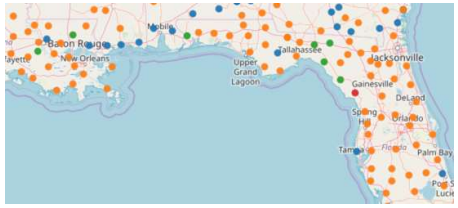
Easy to share!

Output Download

Burundi (Q967)

Murder and nonnegligent manslaughter	6 - point in time: 2002 - applies to people: male - stated in: Sustainable Development

Result sample



Creating Wikidata Satellites: Challenges

- Identify new properties to model satellites
  - Currently done by hand by Knowledge engineers
- Creation of new Qnodes for satellite instances
  - Identified a schema for each satellite
  - Feedback loop to Wikidata
- How to select a “trusty” statement when several values are available?
- Namespace issues
  - Single namespace, or namespace per satellite?
- Inter-satellite linkages

Conclusions

- Tabular data exists in heterogeneous formats
  - Difficult to find, use, augment and share

- WDPlus is a framework to help discover, improve, search, augment, combine and share tabular data

  - WDPlus framework for profiling and enriching datasets

  - T2WML language to generate linked instances from tabular data

  - Encouraging early results on usability

Help us extend WDPlus!

Do you have comments, suggestions or use cases? Contact me at:

dgarijo@isi.edu