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

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

- **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

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

Daniel Garijo and Pedro Szekely. WDPlus: Leveraging Wikidata to Link and Extend Tabular Data. (Sciknow 2019)
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 to check the uploading status
| POST /upload uploaded dataset with metadata
| POST /upload/generateWF+Metadata Generate wikified dataset and its metadata
| POST /upload/merge WithMetadata Use upload dataset with metadata
| POST /upload/uploadWF+Metadata uploaded data and metadata

| Embeddings API | GET /embeddings/fb/[qmode] Fetch the FS 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

Table overview

Entity Linking

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

Easy to share!

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