Towards more consistent import of external data sources to Wikidata
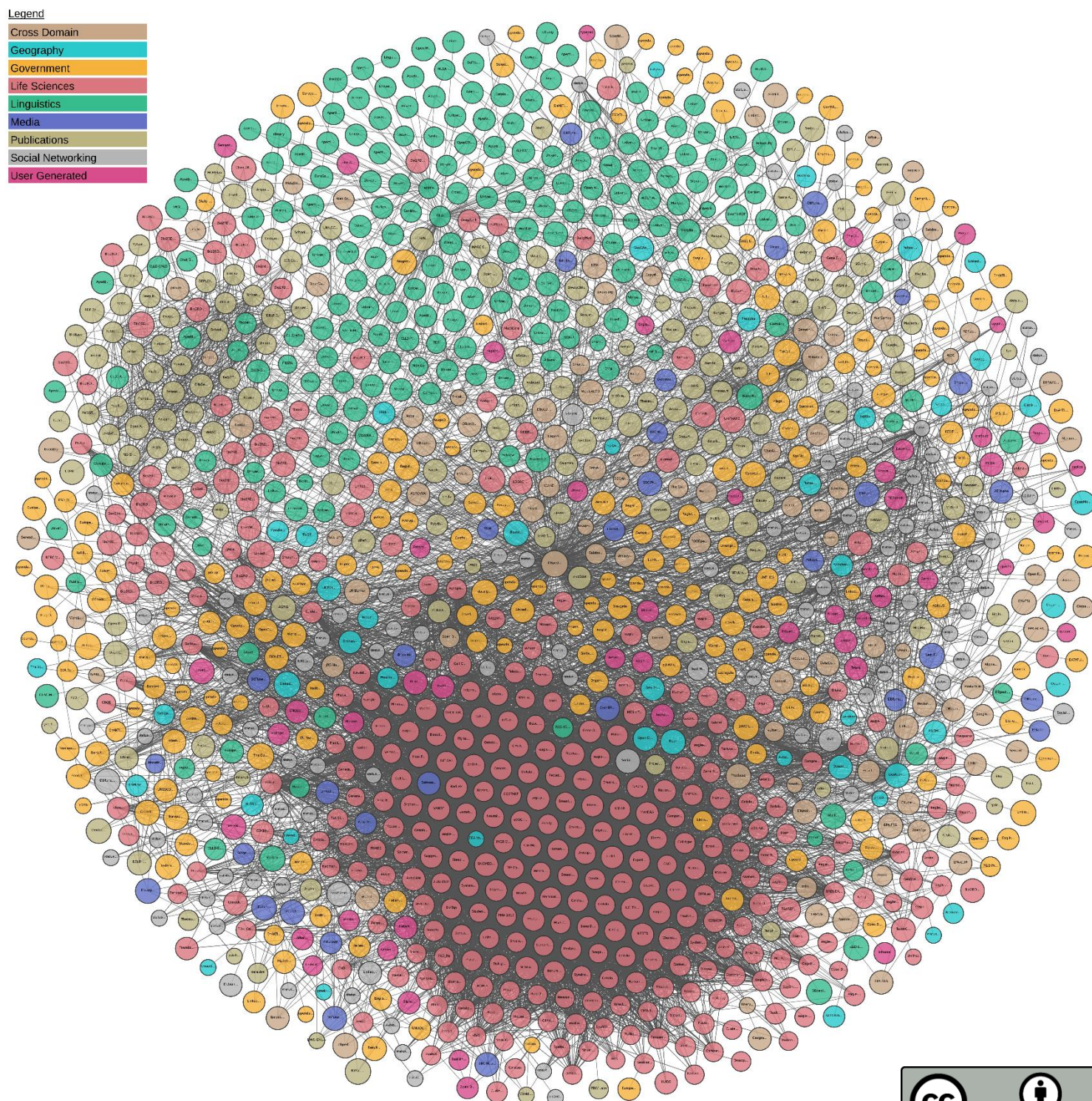
Jakub Klímek and Petr Škoda
Faculty of Mathematics and Physics
Charles University
Prague, Czechia

Project grant
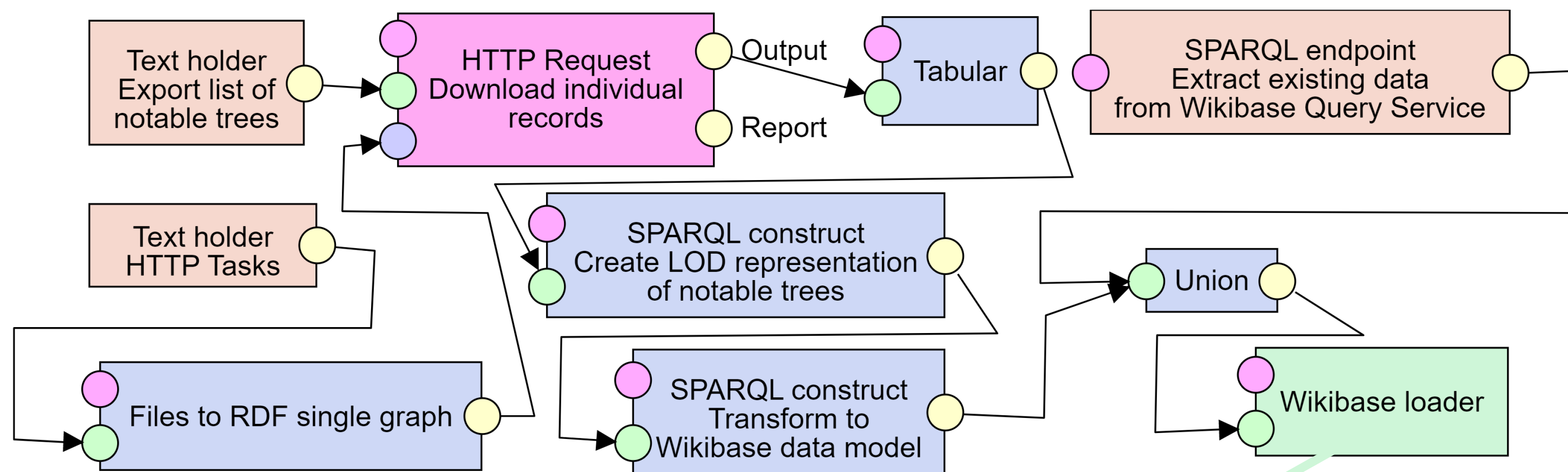Wikidata & ETL

Repeatable
Reusable
Maintainable

Transformations
using RDF in
LinkedPipes ETL

WIKIDATA

WIKIBASE

The Linked Open Data Cloud from lod-cloud.net

Legend:
Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated

LinkedPipes ETL pipeline components:

Text holder
Export list of notable trees

HTTP Request
Download individual records

Output
Report

Tabular

SPARQL endpoint
Extract existing data from Wikibase Query Service

Text holder
HTTP Tasks

SPARQL construct
Create LOD representation of notable trees

Union

Files to RDF single graph

SPARQL construct
Transform to Wikibase data model

Wikibase loader

Wikidata Toolkit wrapped as a component

LinkedPipes ETL data transformation pipeline
1) data from the web downloaded
2) transformed to LOD representation
3) transformed to Wikibase RDF dump representation and changes loaded

LinkedPipes ETL

- Open-source tool for repeatable data transformations
- Currently used to publish and consume Linked Open Data
- Transformations stored as JSON-LD => can be shared (or parts of them) e.g. on GitHub
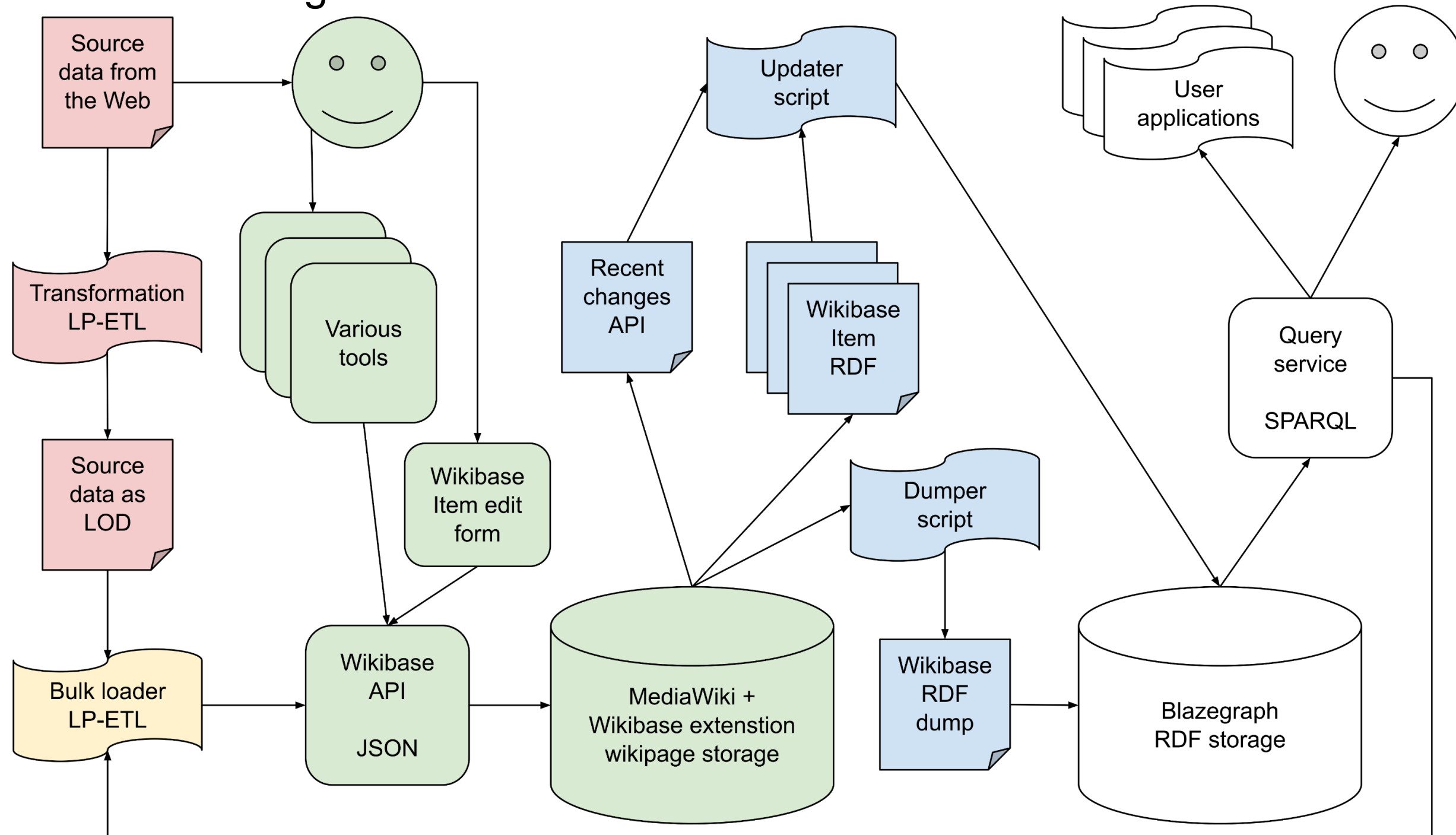
Current Wikidata imports

- Currently mostly ad-hoc scripts using various libraries
- Not easily maintainable by anyone else but the author
- Wikibase (Wikidata) API not easy to use
- Wikibase dump RDF based, but Wikibase API only JSON based.

Source data from the Web

Transformation LP-ETL

Source data as LOD

Bulk loader LP-ETL

Various tools

Wikibase Item edit form

Wikibase API
JSON

Recent changes API

Updater script

Wikibase Item RDF

MediaWiki + Wikibase extension wikipage storage

Dumper script

Wikibase RDF dump

Query service
SPARQL

User applications

Blazegraph RDF storage

Wikibase data flow from import to SPARQL query service

Project start

Analysis done

Demo @ Wikimania

Project end

| 2019 | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |