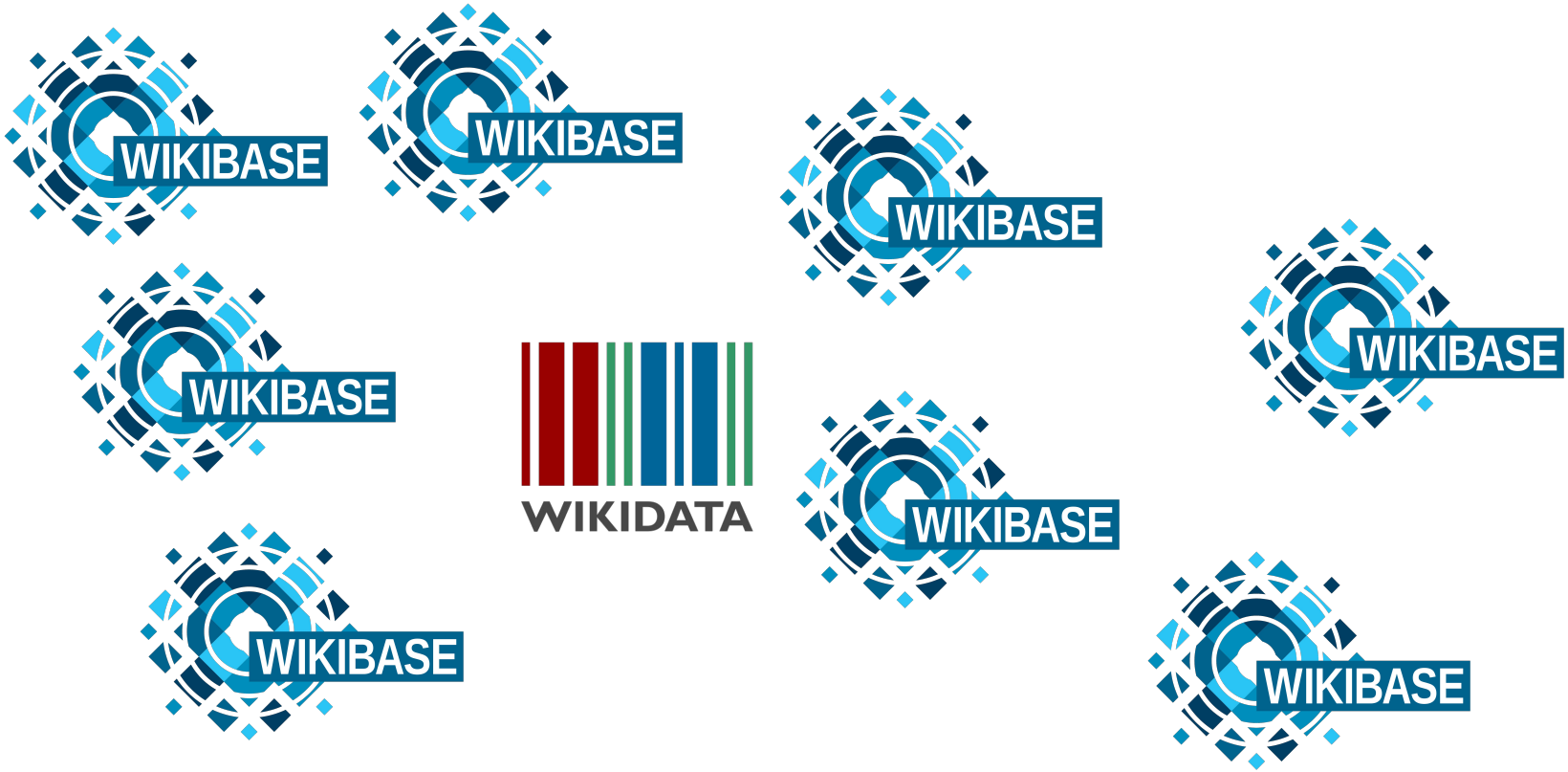
Wikidata & ETL

Jakub Klímek and Petr Škoda

Workshop prerequisites
(for active participation):
RDF, ideally Wikibase RDF dump format

https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format

Wikidata = one of Wikibase instances



Querying Wikibase?

1. Full-text search in Item labels
2. Wikidata Query Service using
   a. RDF dump format
   b. Wikibase ontology (Qs and Ps)
   c. SPARQL query language

Bulk editing Wikibase?

1. Using JSON-based API directly
2. Using a library wrapping the API in custom code
3. Using a semi-automated tool like
   a. OpenRefine
   b. QuickStatements
   c. ...
4. RDF dump format???

Wikidata & ETL: WMF Project Grant

Goal: Allow RDF savvy users to load data to Wikidata using LP-ETL

RDF dump format

ETL = Extract Transform Load

Extract

Get data from somewhere

Transform

Transform the data to target representation

Load

Load the data to a web server, triplestore, Wikibase, etc.

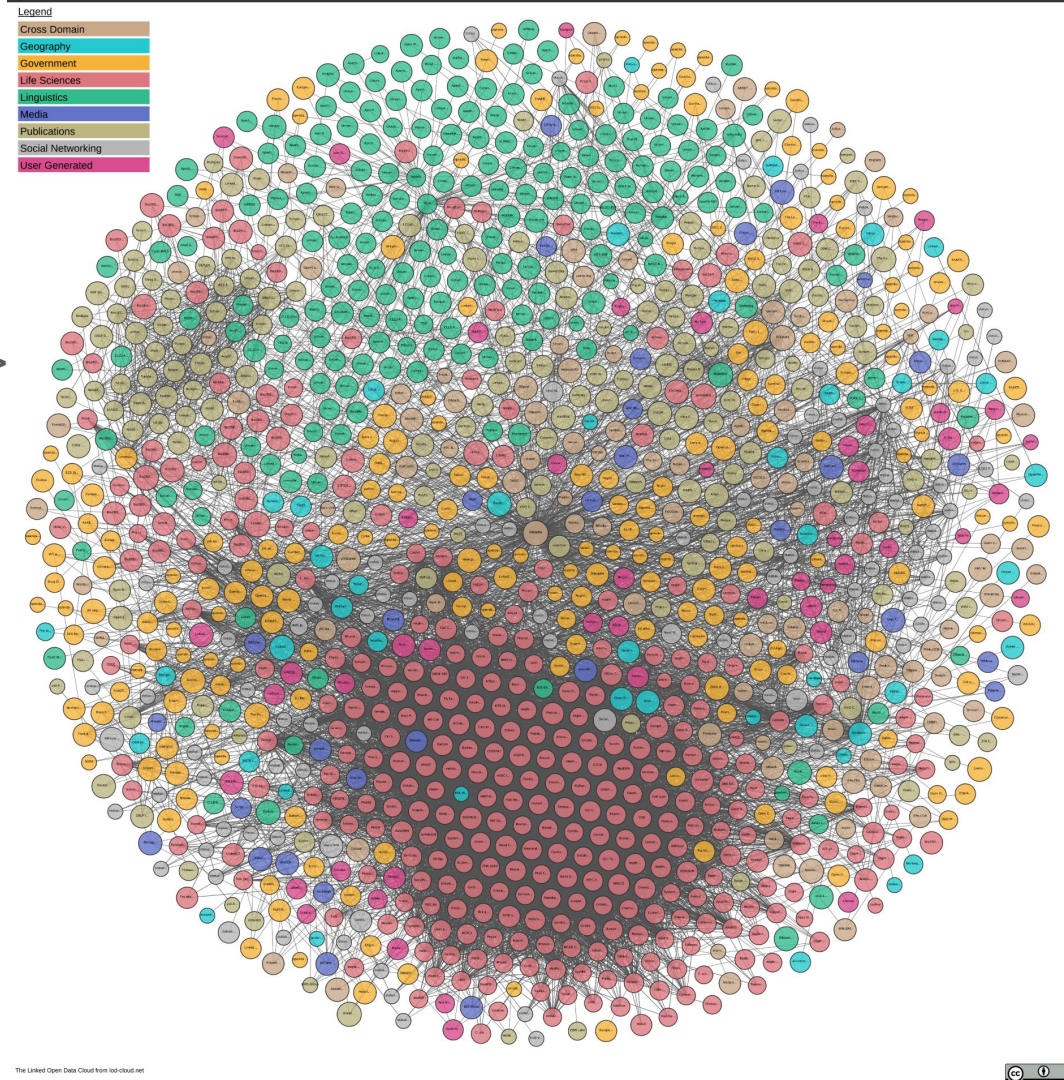
https://en.wikipedia.org/wiki/Extract,_transform,_load

repeatable pipeline

LinkedPipes ETL (LP-ETL)

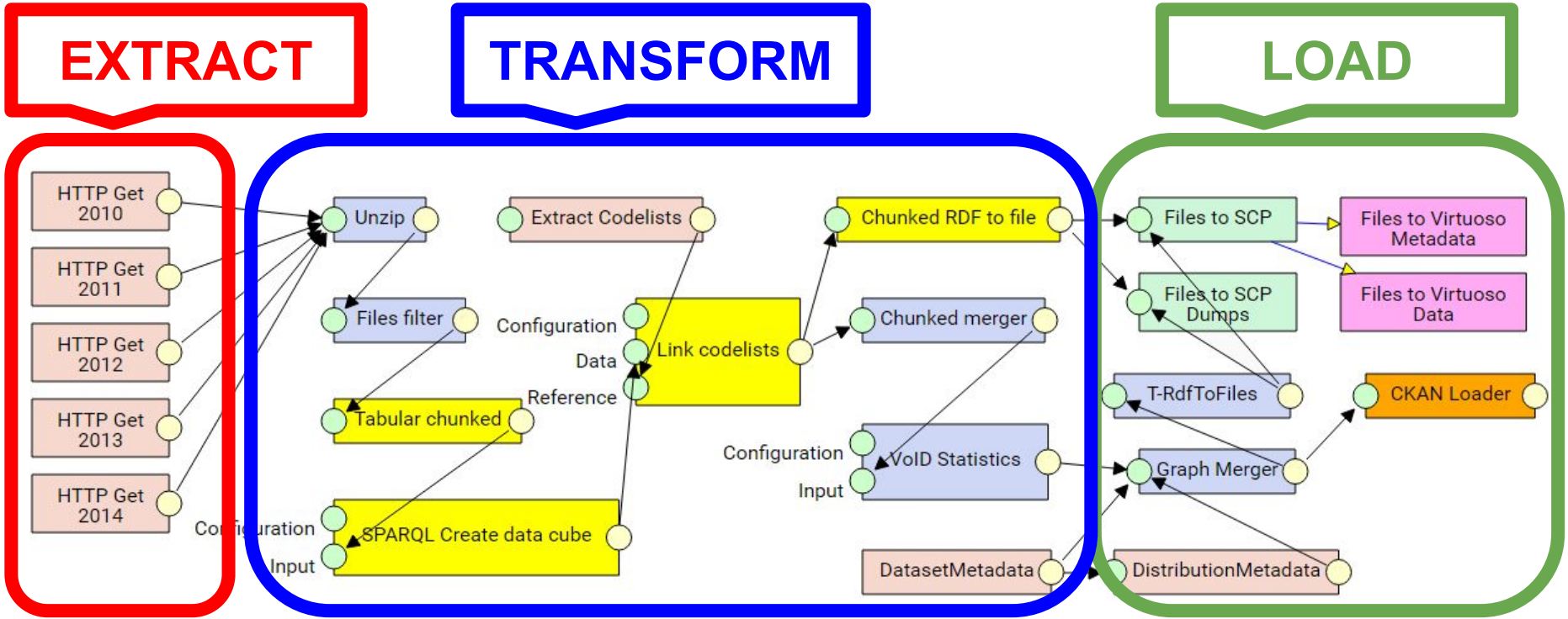
- used with Linked Open Data >>
- open-source ETL tool
- consumption and publication
- RDF + SPARQL oriented
- reusable components
  - XML, JSON, CSV, RDF, ...
- repeatable data pipelines

https://demo.etl.linkedpipes.com



ETL = Extract Transform Load

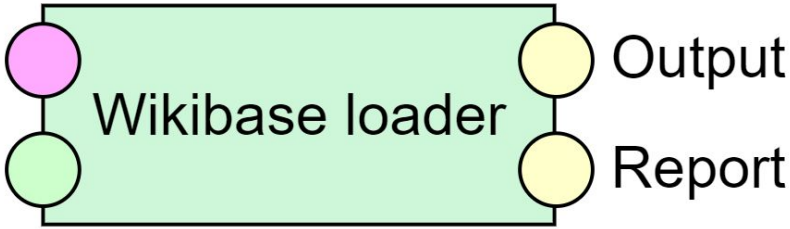
components in a **pipeline**



Wikidata & ETL

Goal: Allow RDF savvy users to load data to Wikidata using LP-ETL

- Technical solution: LP-ETL component: Wikibase Loader
- Inputs:
  - Data in Wikibase RDF dump format
  - Wikibase api.php URL
  - Wikibase ontology prefix
  - Credentials - username and password



- Results in data loaded into (updated in) Wikibase
- In a repeatable pipeline
  - when source data changes and the pipeline runs, Wikibase is updated as well
- Using reusable components - better than custom code

What is happening in the pipeline?
Updating an item/statement = using its URL

```
wd:Q2175 a wikibase:Item ;
    rdfs:label "Test item"@en ;
    p:P8 wds:Q2175-F9E61242-E432-4DC3-873E-8A004706EBA7 .

wds:Q2175-F9E61242-E432-4DC3-873E-8A004706EBA7 a wikibase:Statement ;
    psv:P8 <urn:value1> .

<urn:value1> a wikibase:QuantityValue ;
     wikibase:quantityAmount "5"^^xsd:decimal ;
     wikibase:quantityUpperBound "10"^^xsd:decimal ;
     wikibase:quantityLowerBound "5"^^xsd:decimal ;
     wikibase:quantityUnit <https://wikibase.opendata.cz/entity/Q2106> .
```

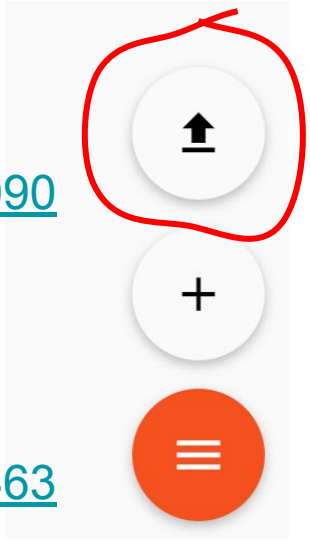
What is happening in the pipeline?
Creating an item/statement = tagging it as “new”

```
@prefix new: <http://plugins.linkedpipes.com/ontology/l-wikibase#New> .
<urn:NewItem> a wikibase:Item, new: ;
    rdfs:label "qualifierTest item"@en ;
    p:P12 <urn:NewStatement> .

<urn:NewStatement> a wikibase:Statement, new: ;
    psv:P12 <urn:value1> .

<urn:value1> a wikibase:GlobecoordinateValue ;
     wikibase:geoLatitude "12.933333333333"^^xsd:double ;
     wikibase:geoLongitude "35.3"^^xsd:double ;
     wikibase:geoPrecision "0.000277778"^^xsd:double ;
     wikibase:geoGlobe <http://www.wikidata.org/entity/Q2> .
```

For you to try NOW!

Use a pre-prepared demo pipeline creating a sample item in our Wikibase instance

1. LP-ETL demo instance: https://demo.etl.linkedpipes.com
2. https://wikibase.opendata.cz - see Recent changes (enable bots)
3. Register yourself or use our credentials - this is really a sandbox
4. Import pipeline https://demo.etl.linkedpipes.com/resources/pipelines/1565946968090
5. Rename the important pipeline (to know which one is yours)



Demo pipeline editing the Wikidata Sandbox Item

1. Import pipeline https://demo.etl.linkedpipes.com/resources/pipelines/1565948341463
2. Use our bot credentials
   a. the LP-ETL demo instance is public - do not leak your Wikidata credentials
3. (There is a bug - the pipeline will fail, but the edit is done anyway)