FUJITSU

shaping tomorrow with you

JIST2018 Tutorial
Practical Use of a Knowledge Graph with Case Studies using Semantic Web Publishing Tools

2018.11.26

Fujitsu Laboratories LTD.
Kenji Kobayashi

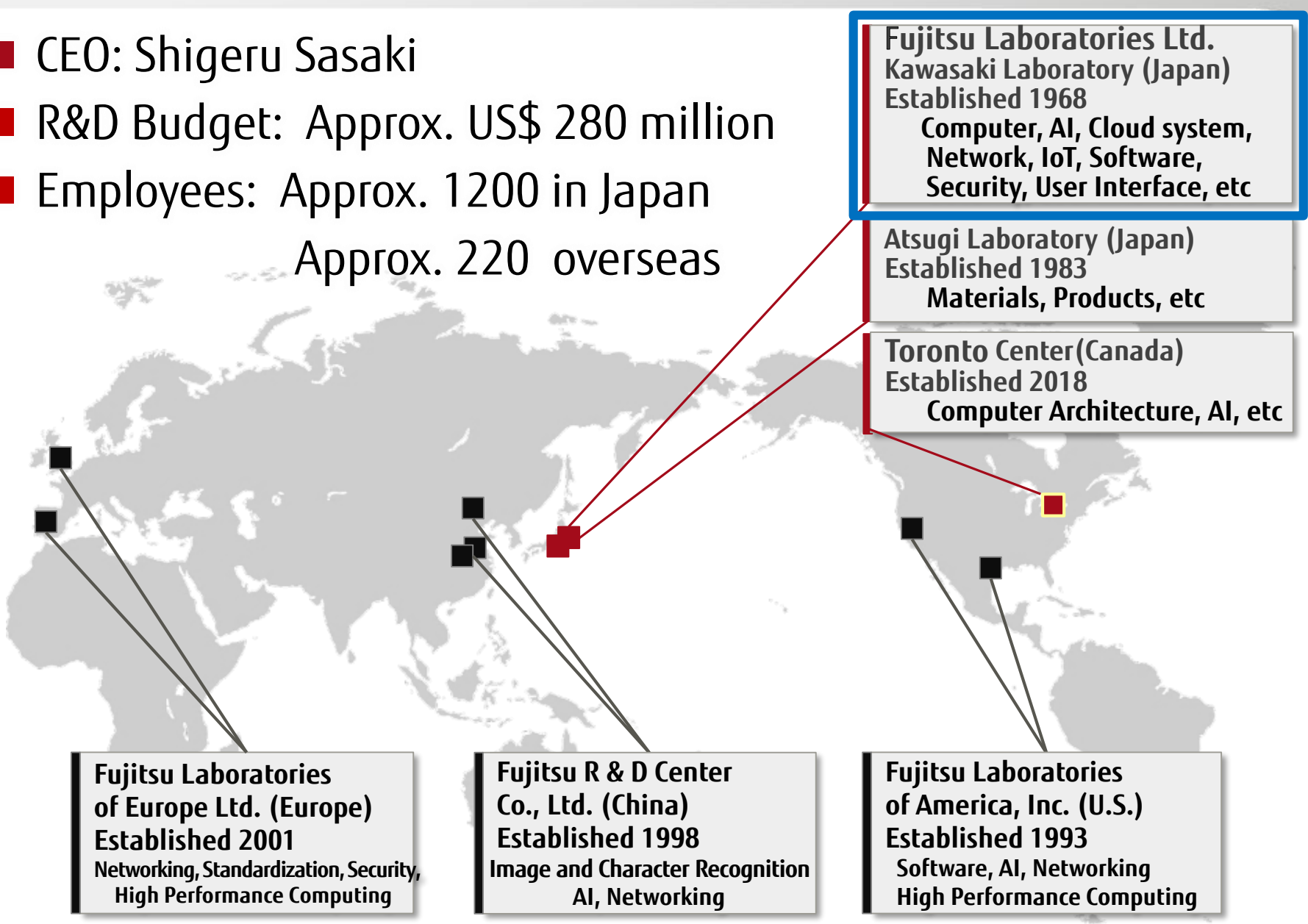
Introduction : Schedule

FUJITSU

Morning	Introduction to Linked Data tools	9:00-10:00 (60 mins)	Presentation
		10:00-10:10 (10 mins)	Break
		10:10-10:50 (40 mins)	Presentation
		10:50-11:10 (20 mins)	Use Case Studies
Afternoon	Hands-on	13:00-13:10 (10 mins)	Introduction
		13:10-14:55 (95 mins)	Trial 1: Let's visualize corporate data Trial 2: Let's develop a financial analysis application
		14:55-15:00 (5 mins)	Closing

Introduction: Fujitsu Laboratories LTD.

- CEO: Shigeru Sasaki
- R&D Budget: Approx. US$ 280 million
- Employees: Approx. 1200 in Japan
  Approx. 220 overseas



Introduction : Tutors

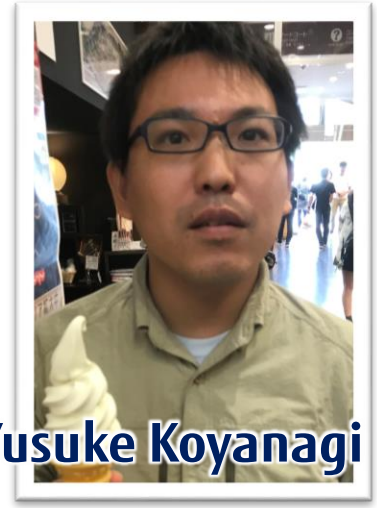
FUJITSU


Kenji Kobayashi
(Presenter)

Hiroaki Morikawa
(Presenter)

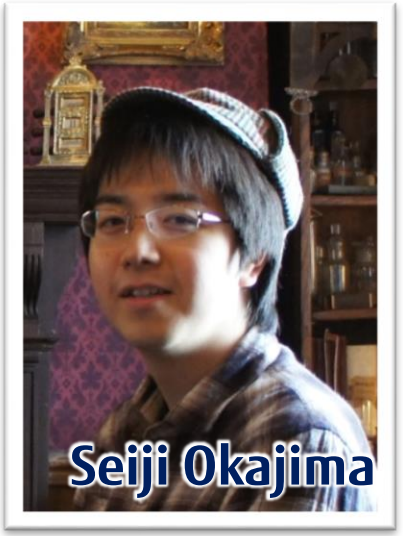

Takanori Ugai


Fumihito Nishino


Yusuke Koyanagi


Bin Piao

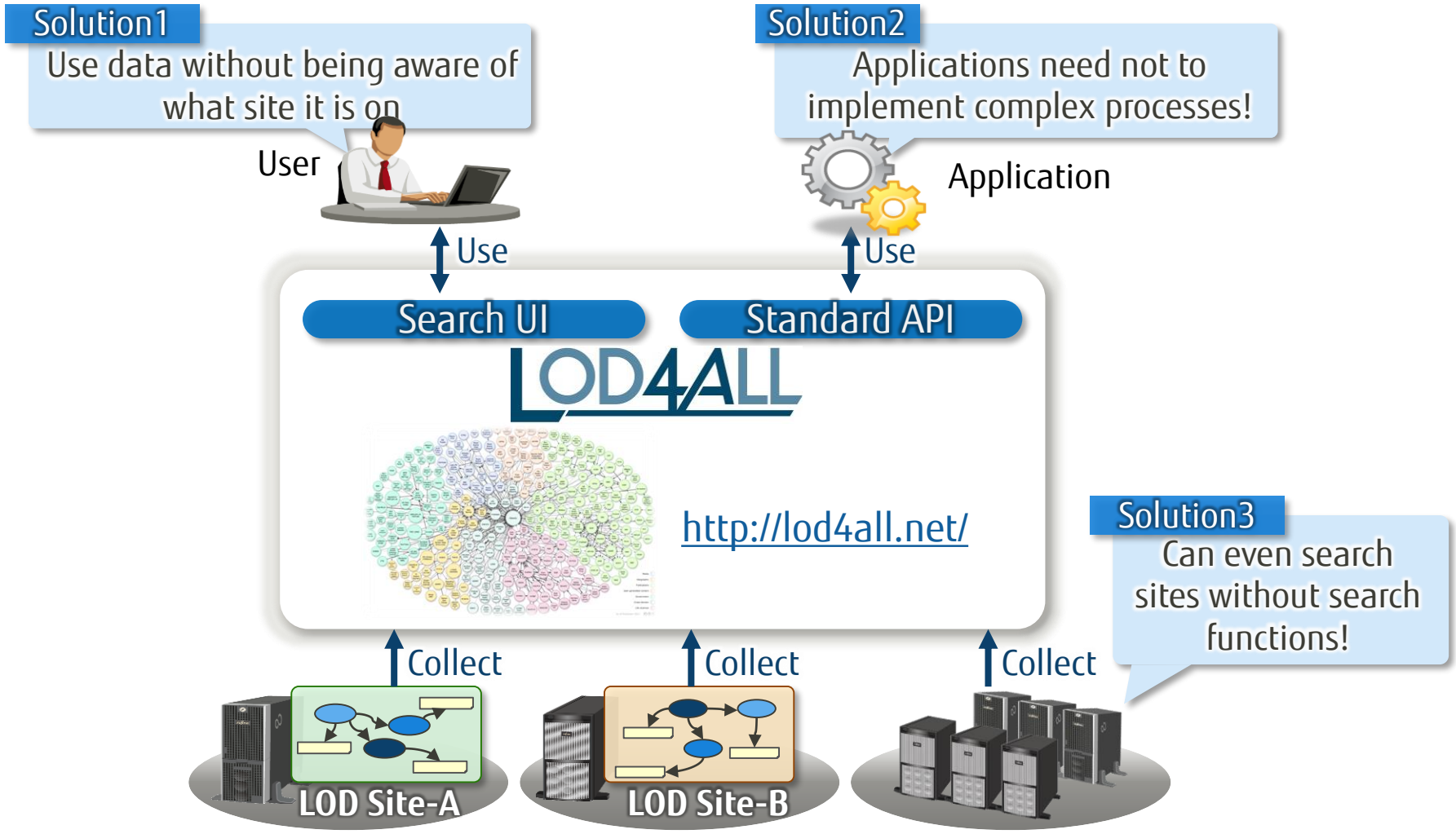

Kyomoto Matsushita


Seiji Okajima

- Department: AI Laboratory
- Research Focus: Semantic Web, Linked Data (LD), Open Data

Introduction: LOD Search Engine

- Crawl LOD from around the world (several tens of billion data items), and provide a high-speed search service



Introduction : Our Research

FUJITSU

- Publications
  - Manuel Peña Muñoz, Alejandro Llaves and Terunobu Kume. Populating the FLE Financial Knowledge Graph Proceedings of the 17th International Semantic Web Conference
  - Hiroaki Morikawa,H., Nishino,F: Design and Implementation of a Standpoint-based Linked Data Visualization Approach n Proc. of the JIST 2018 Poster and Demonstrations Track. (2018).
  - Villazon Terrazas, Boris & Garcia-Santa, Nuria & San Miguel, Beatriz & del Rey-Mejías, Angel & Muria Tarazon, Juan Carlos & Seara, Germán & Reneses, Blanca & de la Torre, Victor. (2018). Fujitsu HIKARI, a Healthcare Decision Support System based on Biomedical Knowledge. International Journal of Privacy and Health Information Management. 6. 26-49. 10.4018/IJPHIM.2018070103.
  - Abe, S., Mitsuishi, Y., Tago, S., Igata, N., Okajima, S., Morikawa, H., Nishino, F.: Linked Corporations Data in Japan. In Proc. of the ISWC 2016 Poster and Demonstrations Track. (2016).
  - Naseer, A., Kume, T., Izu, T., Igata, N.: LOD for All: Unlocking infinite opportunities. In: The Semantic Web Challenge 2014, The 13th International Semantic Web Conference (2014)
  - Kobashi, H., Carvalho, N., Hu B., Saeki, T.: Cerise: an RDF store with adaptive data reallocation. In Proc. of the 13th Workshop on Adaptive and Reflective Middleware Article No. 1. (2014)
  - Carlos Buil-Aranda, Aidan Hogan, Jurgen Umbrich, and Pierre-Yves Vandenbussche. SPARQL Web-Querying Infrastructure: Ready for Action? Proceedings of the 12th International Semantic Web Conference, (Oct. 2013). *Best Paper Award

Introduction : Our Research (cont.)

FUJITSU

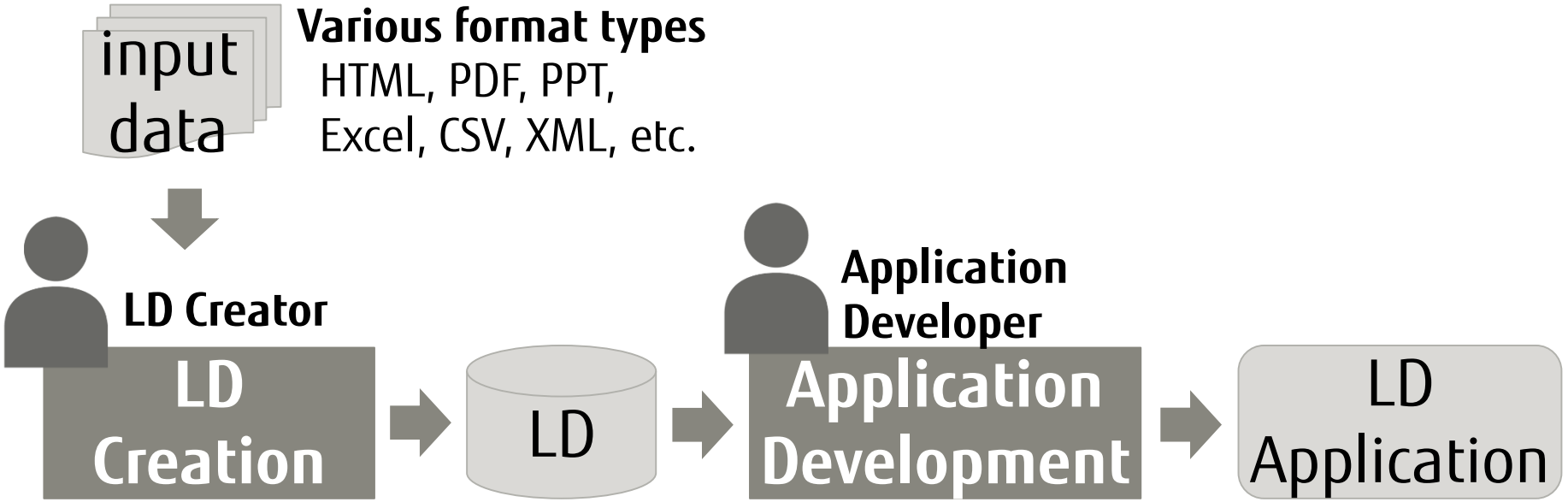
- Publications (cont.)
  - Shohei Yamane, Takanori Ugai, Conversion of Physical Quantity and its Application, Proceedings of the 17th International Semantic Web Conference Poster and Demo,
  - Lu Fang, Qingliang Miao and Yao Meng. DBpedia Entity Type Inference Using Categories
  - Bo Hu, Aisha Naseer, Eduarda M. Rodrigues, Pierre-Yves Vandenbussche, and Masatomo Goto (2013). Linked Data for Financial Reporting. Fourth International Workshop on Consuming Linked Data, collocated with the 12th International Semantic Web Conference, (Oct. 2013).
  - Pierre-Yves Vandenbussche. Empower Flexibility via Big Data: Unleashing Constraints in Financial Analytics. Proceedings of the 26th XBRL conference, (Apr. 2013).
  - Pierre-Yves Vandenbussche, Carlos Buil-Aranda, Aidan Hogan, and Jurgen Umbrich. Monitoring SPARQL Endpoint Status. Proceedings of the 12th International Semantic Web Conference Poster and Demo, (Oct. 2013).
- Award
  - **Won Corporation Information Award** in RESAS Application Contest Organized by METI(Ministry of Economy, Trade and Industry)

Introduction : Contents

FUJITSU

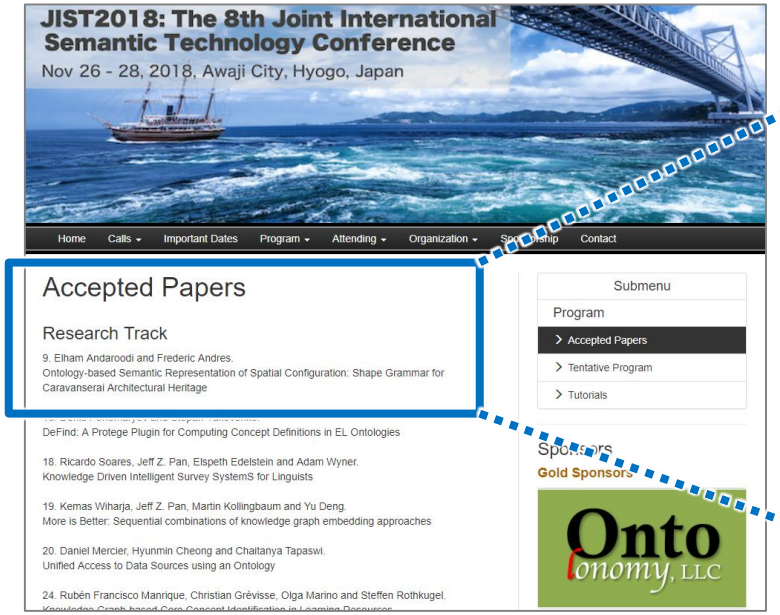
This tutorial covers useful tools*, from Linked Data creation to application development using Linked Data.

*used by us, and license free

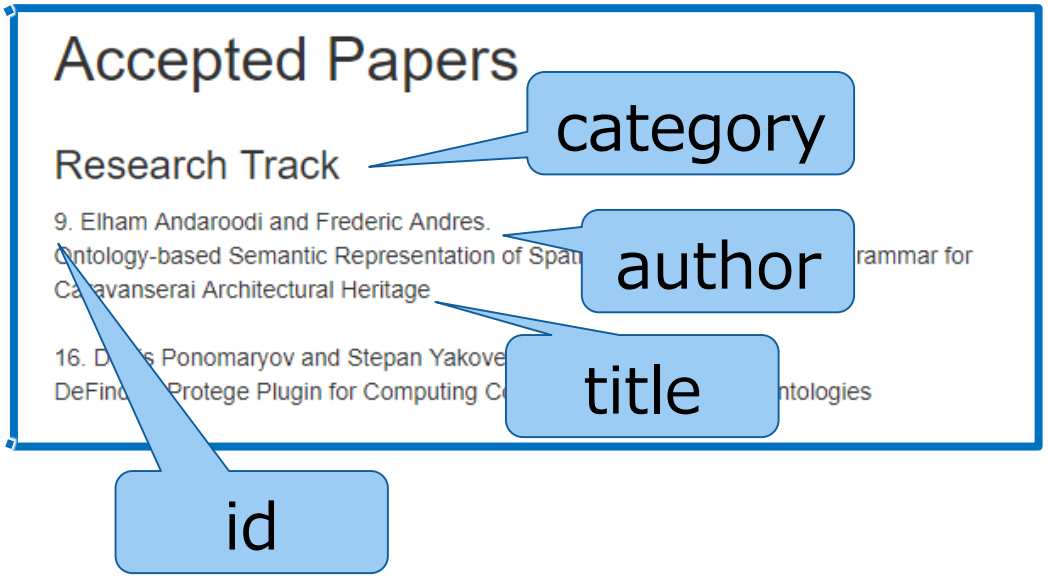


Introduction : Sample Input Data

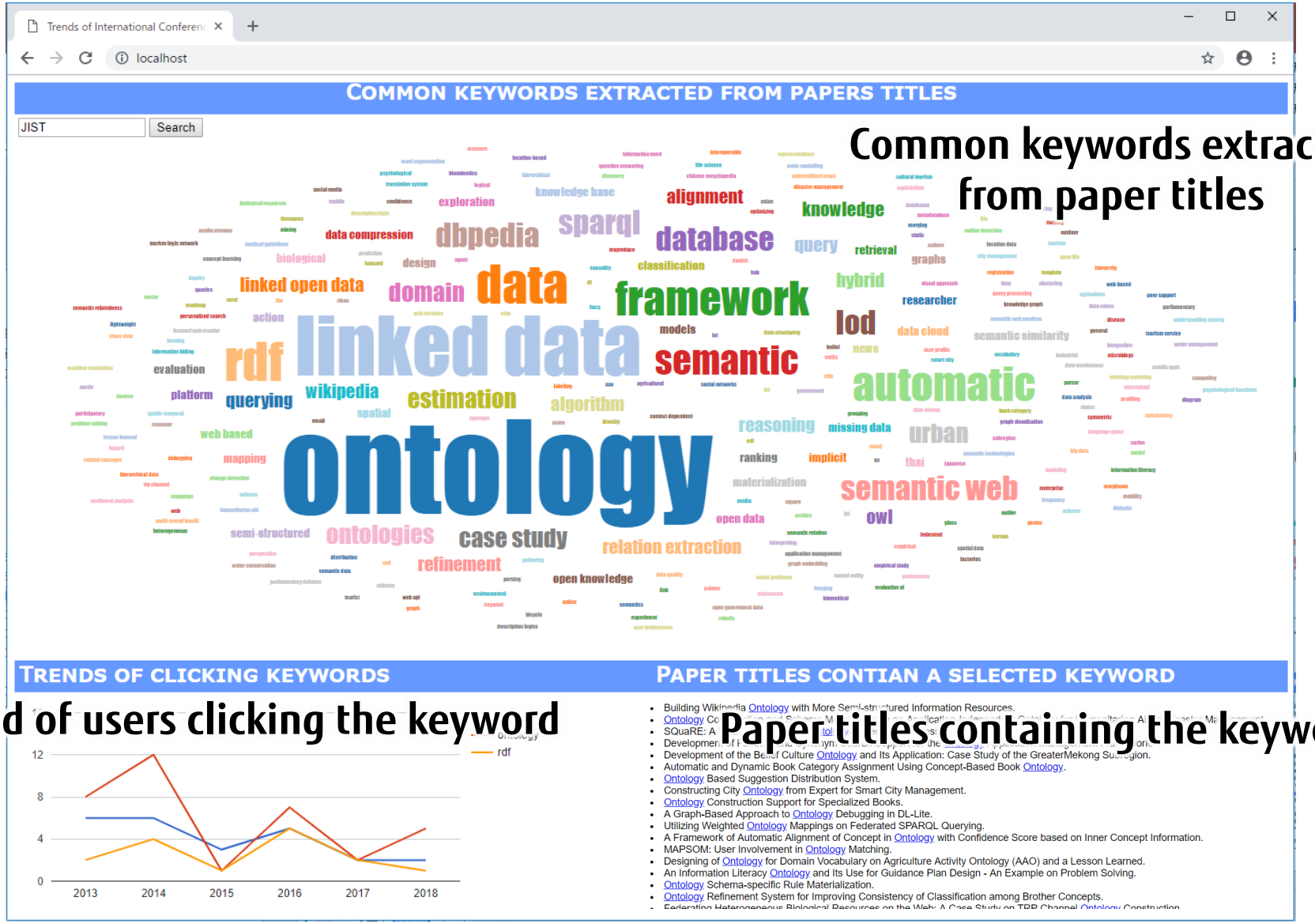
- *Accepted Paper* page on the JIST2018 website
  - HTML format
  - Each paper has {id, author, title, category}



JIST2018 website



Introduction : Sample LD Application



Common keywords extracted from paper titles

Trend of users clicking the keyword

Paper titles containing the keyword

Outline

FUJITSU

input data

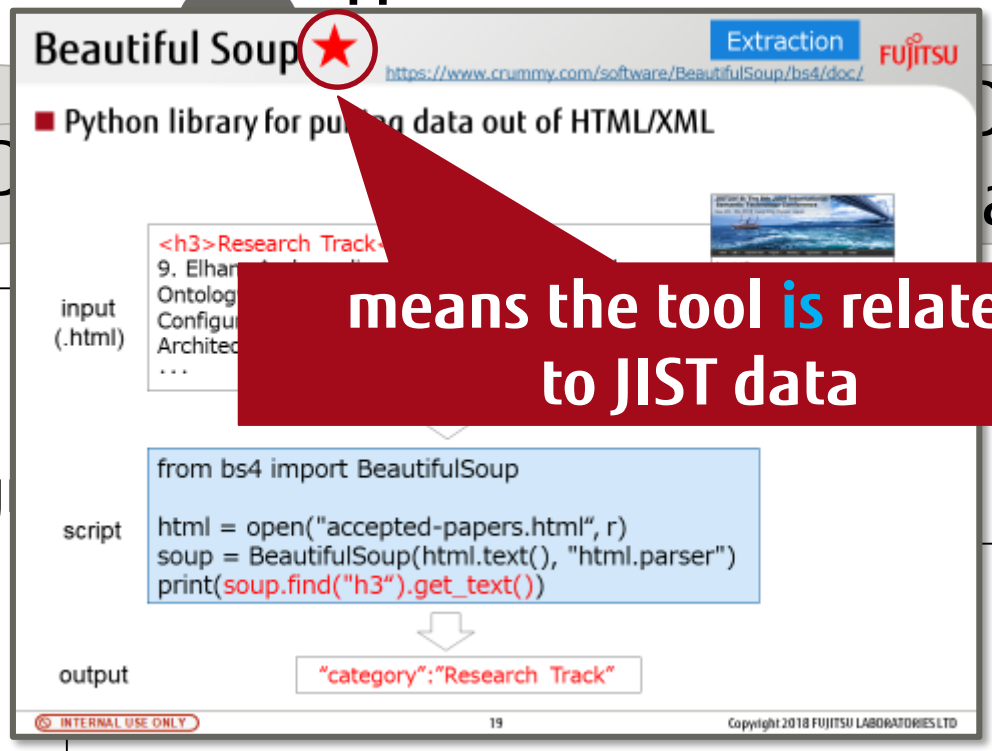
Various format types
HTML, PDF, PPT,
Excel, CSV, XML, etc.

LD Creator

LD Creation

1. Extraction
2. Cleansing
3. RDF Schema Desig
4. Conversion to RDF
5. Linking
6. Validation
7. DB Construction

Application

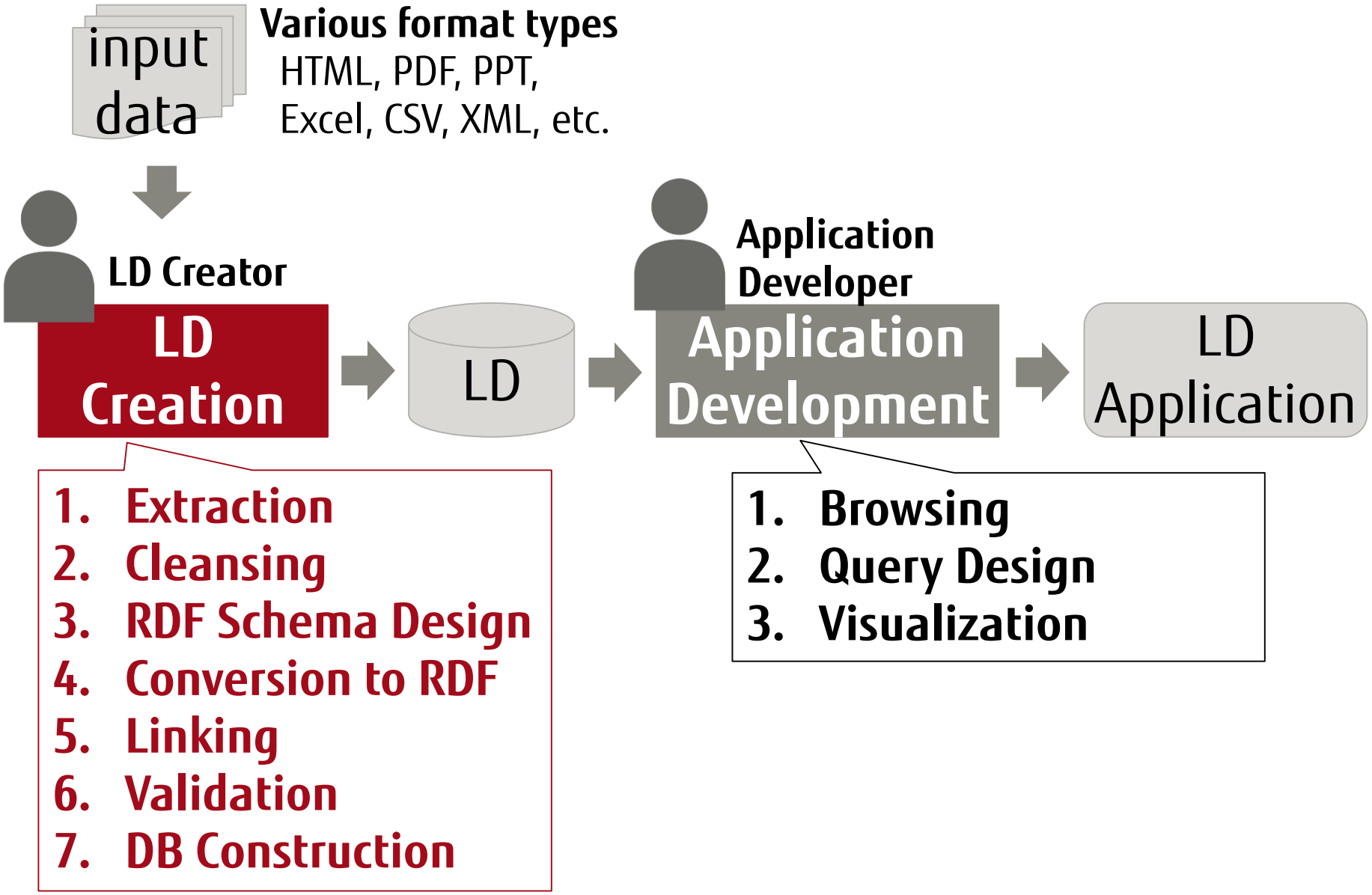


means the tool is related to JIST data

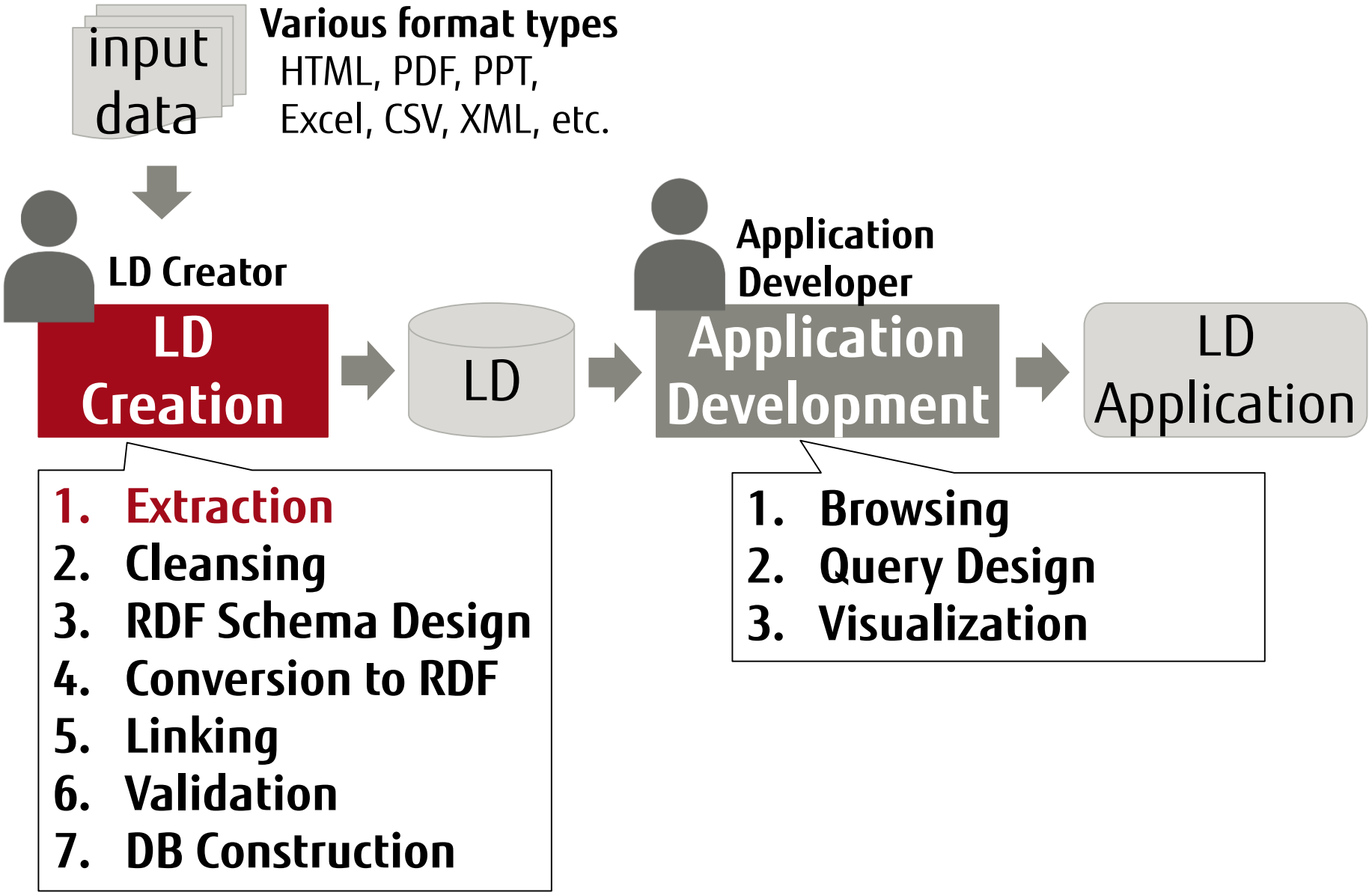
FUJITSU

1. LD Creation

Outline



Outline

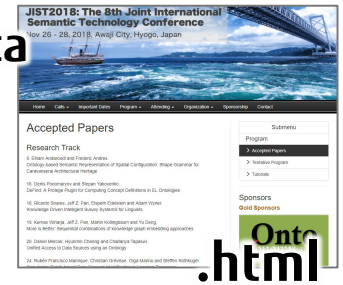


Extraction

FUJITSU

Extract useful data from various data formats

Ex. JIST data



.html

Extraction

```
{"paperId":"id_x",
 "author":"author_x" ,
 "title":"title_x"}
```

.json

- From HTML documents
  - Beautiful Soup★
- From PDF documents
  - pdfminer, tabula, pdfx
- From PowerPoint documents
  - python-pptx

Beautiful Soup ★

Extraction(1/6)

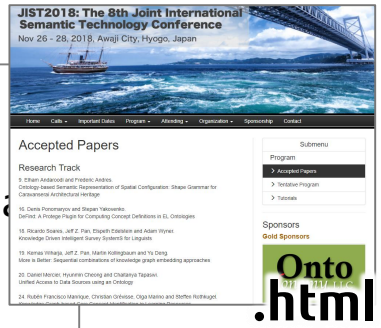
https://www.crummy.com/software/BeautifulSoup/bs4/doc/

FUJITSU

- Python library for pulling data out of HTML/XML

input (.html)

```
<h3>Research Track</h3>
9. Elham Andaroodi and Frederic Andres. <br>
Ontology-based Semantic Representation of Spatia
Configuration: Shape Grammar for Caravanserai
Architectural Heritage<br>
...
```



script

```
from bs4 import BeautifulSoup

html = open("accepted-papers.html“, r)
soup = BeautifulSoup(html.text(), "html.parser")
print(soup.find("h3“).get_text())
```

output

```
“category”:”Research Track”
```

pdfminer

https://www.unixuser.org/~euske/python/pdfminer/

FUJITSU

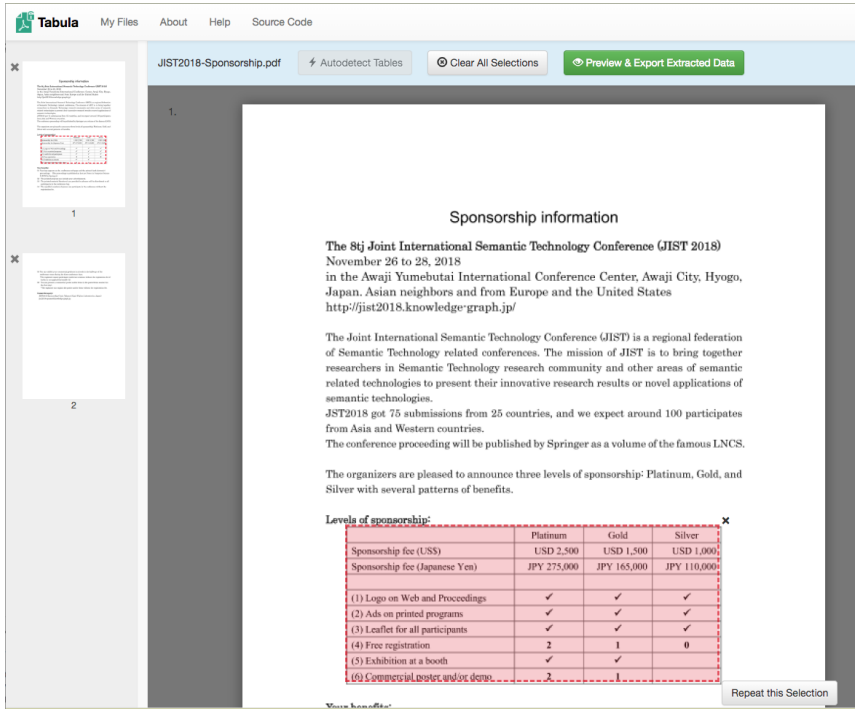
- Command-line tool for extracting information from PDF documents

```
$ pdf2txt.py -t html p177-constantin.pdf
<html><head>
<meta http-equiv="Content-Type" content="text/html;
charset=utf-8">
</head><body>
<span style="position:absolute; border: gray 1px solid;
left:0px; top:50px; width:612px; height:792px;"></span>
<div style="position:absolute; top:50px;"><a
name="1">Page 1</a></div>
<div style="position:absolute; border: textbox 1px
solid; writing-mode:lr-tb; left:90px; top:116px;
width:429px; height:22px;">
...
```

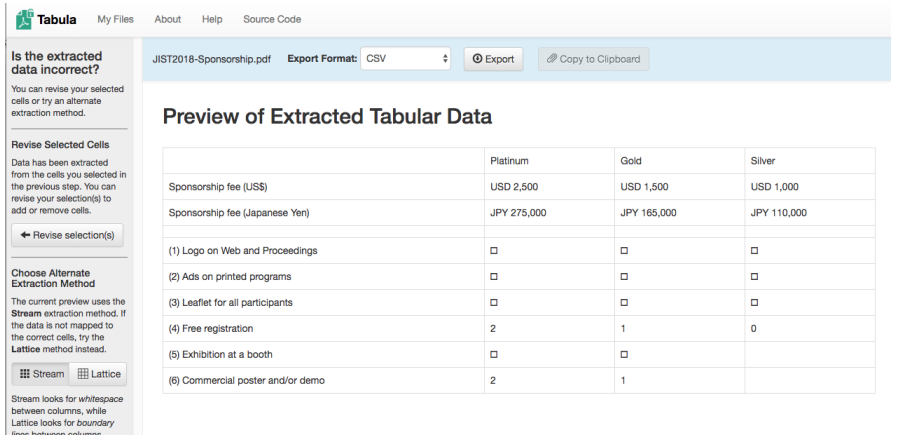
Tabula

https://github.com/tabulapdf/tabula/releases/tag/v1.2.1

- Converts tables in PDF files to CSV
  - It generates XML with coordinates as intermediate data, and recognizes words, cell borders, and tables
  - WebUI



Automatic recognition tables in PDF



	Platinum	Gold	Silver
Sponsorship fee (US$)	USD 2,500	USD 1,500	USD 1,000
Sponsorship fee (Japanese Yen)	JPY 275,000	JPY 165,000	JPY 110,000
(1) Logo on Web and Proceedings	□	□	□
(2) Ads on printed programs	□	□	□
(3) Leaflet for all participants	□	□	□
(4) Free registration	2	1	0
(5) Exhibition at a booth	□	□	
(6) Commercial poster and/or demo	2	1	

Extraction result

Tabula-py

https://github.com/chezou/tabula-py

FUJITSU

- Python wrapper of Tabula
  - It recognizes tables in PDF automatically as well as Tabula (Default)
    - manual setting is also possible

In the case of setting a fixed area

Set a fixed area in PDF

Ex. Using the Mac Preview function

http://docs.whirlpool.eu/_doc/Cookbook_Mwo_only_501912000448EN.pdf

Code

```
from tabula import read_pdf

df = read_pdf("recipes2.pdf", guess=False,
              area=(50.0, 200.0, 82.0, 548.0),
              pages='all')
```

set

Extraction result

Cheese and Mushroom Toast
Asparagus and Cheese Soup
Tomato Soup
Mushroom Soup
Mushroom and Saffron Risotto
Pork with Onions and Peppers
Jacket Potatoes
Vegetable Ratatouille
Tomato Sauce

- Converts scholarly papers (in English) to HTML/XML
  - Web Service

PDF

PDFX: Fully-automated PDF-to-XML Conversion of Scientific Literature

Alexandru Constantin aconstantin@cs.man.ac.uk
Steve Pettifer steve.pettifer@cs.man.ac.uk
Andrei Voronkov voronkov@cs.man.ac.uk

School of Computer Science
The University of Manchester, United Kingdom
Oxford Road, M13 9PL

ABSTRACT

PDFX is a rule-based system designed to reconstruct the logical structure of scholarly articles in PDF form, regardless of their formatting style. The system's output is an XML document that describes the input article's logical structure in terms of title, sections, tables, references, etc. and also links it to geometrical typesetting markers in the original PDF, such as paragraph and column breaks. The key aspect of the presented approach is that the rule set used relies on relative parameters derived from font and layout specifics of each article, rather than on a template-matching paradigm. The system thus obviates the need for domain- or layout-specific tuning or prior training, exploiting only typographical conventions inherent in scientific literature. Evaluated against a significantly varied corpus of articles from nearly 2000 different journals, PDFX gives a 77.45 F1 measure for top-level heading identification and 74.03 for extracting individual bibliographic items. The service is freely available for use at http://pdfx.cs.man.ac.uk/.

Categories and Subject Descriptors

I.7.5 [Document and Text Processing]: Document Capture—*document analysis*; I.2.7 [Artificial Intelligence]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Design

Keywords

document structure analysis; PDF conversion; logical structure recovery; PDFX

1. INTRODUCTION

The recent increase in volume of the global research output has given rise to numerous initiatives that focus on automatic document processing. The predominant goal of these approaches has been to reduce the search space for potentially relevant information through means such as intuitive indexing and retrieval, document summarisation or discourse annotation. The added value brought by such services can be quite significant when considering the expeditious publication rate that certain fields of study enjoy, for instance biomedicine and the life sciences. Many such tools, however, work exclusively on plain-text and not camera-ready, typeset publications. This makes their performance dependent on data sources containing noise-free, accurate reconstructions of article narratives. Automating this pre-processing step requires programmatic access to the typographical layout of elements on page as well as to their logical/rhetorical function within the article. For this reason, analysis tools of scientific text often choose to couple themselves to data stores that have human-curated semi-structured representations of articles readily available, such as PMC [3], Scopus [7], DBLP [12] or arXiv [1]. This dependency deters the tools' widespread adoption because much of the information sought by researchers is made available solely within PDF publications with no alternative representation. Without the means to expand their reach to this highly popular format, many promising natural language processing and text mining solutions remain either undiscovered or of little use to potential users.

Notable previous efforts targeting the recovery of structure from the PDF are given in Table 1 along with their capabilities. Except for the machine learning solution SectLabel [13], the tools focus on geometrical analysis and either do not handle or are in their preliminary phases of logical structure recognition. In this paper we present a novel system called PDFX that focuses on logical structure, but handles its geometrical baseline as well. It aims to identify, extract and link these two structures together in order to facilitate an enhanced level of interaction with an article's contents. The method employed is rule-based, iterative and unrestricted with respect to the set of formatting templates that input articles need to adhere to. The only requirement is that they be full-text natively typeset PDF publications, as opposed to PDF images such as scans of paper documents. The 18 logical element types that the system can currently identify (listed in Table 2) cover the principal parts of a typical research article. They are ultimately stored in an XML file with a tag hierarchy that closely follows the JATS standard[1]. The semi-structured nature of the XML serves as a convenient, quick-access route to any of the articles's components.

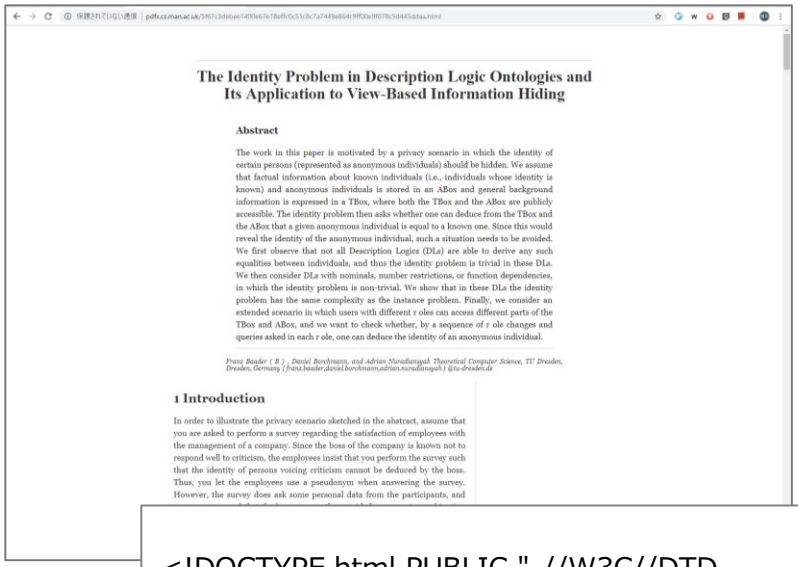
2. SYSTEM AND METHODS

PDFX carries out a two-stage process in order to address structure recovery. The first stage constructs a geometrical model of the article's contents to determine the spatial organisation of textual and graphical units on page. The second stage draws upon the

[1]ANSI/NISO Z39.96-2012 standard. JATS: Journal Article Tag Suite - http://jats.nlm.org/ - formalised from the NLM Archiving and Interchange Tag Suite

pdfx

HTML



The Identity Problem in Description Logic Ontologies and Its Application to View-Based Information Hiding

Abstract

The work in this paper is motivated by a privacy scenario in which the identity of certain persons (represented as anonymous individuals) should be hidden. We assume that factual information about known individuals (i.e., individuals whose identity is known) and anonymous individuals is stored in an ABox and general background information is expressed in a TBox, where both the TBox and the ABox are publicly accessible. The identity problem then asks whether one can deduce from the TBox and the ABox that a given anonymous individual is equal to a known one. Since this would reveal the identity of the anonymous individual, such a situation needs to be avoided. We first observe that not all Description Logics (DLs) are able to derive any such equalities between individuals, and thus the identity problem is trivial in these DLs. We then consider DLs with nominals, number restrictions, or function dependencies, in which the identity problem is non-trivial. We show that in these DLs the identity problem has the same complexity as the instance problem. Finally, we consider an extended scenario in which users with different r oles can access different parts of the TBox and ABox, and we want to check whether, by a sequence of r ole changes and queries asked in each r ole, one can deduce the identity of an anonymous individual.

Franz Baader (B) , Daniel Borchmann, and Adrian Nuradiansyah Theoretical Computer Science, TU Dresden, Dresden, Germany {franz.baader,daniel.borchmann,adrian.nuradiansyah}@tu-dresden.de

1 Introduction

In order to illustrate the privacy scenario sketched in the abstract, assume that you are asked to perform a survey regarding the satisfaction of employees with the management of a company. Since the boss of the company is known not to respond well to criticism, the employees insist that you perform the survey such that the identity of persons voicing criticism cannot be deduced by the boss. Thus, you let the employees use a pseudonym when answering the survey. However, the survey does ask some personal data from the participants, and

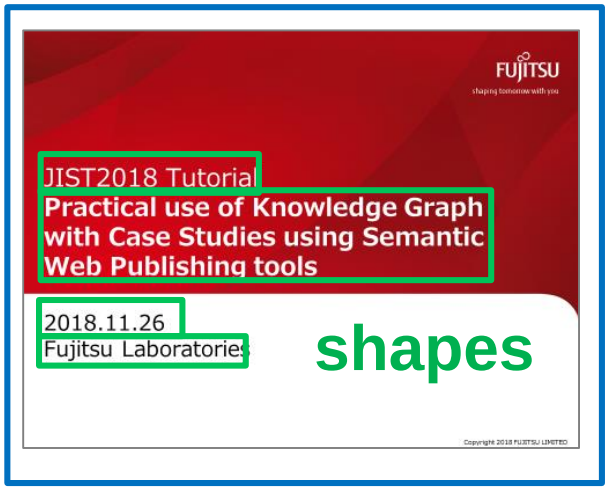
```
<!DOCTYPE html PUBLIC "-//W3C//DTD
XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-
transitional.dtd">
<html
xmlns="http://www.w3.org/1999/xhtml"><h
ead><meta http-equiv="Content-Type"
content="text/html; charset=UTF-8">
                                    <title>pdfx -
HTML for "The Identity Problem ..."</title>
```

python-pptx

https://python-pptx.readthedocs.io/en/latest/

FUJITSU

- Python library for extracting information from PowerPoint (.pptx) documents
  - It extracts not only text boxes, but also graphs, tables, and so on, with page numbers and coordinates



python-pptx

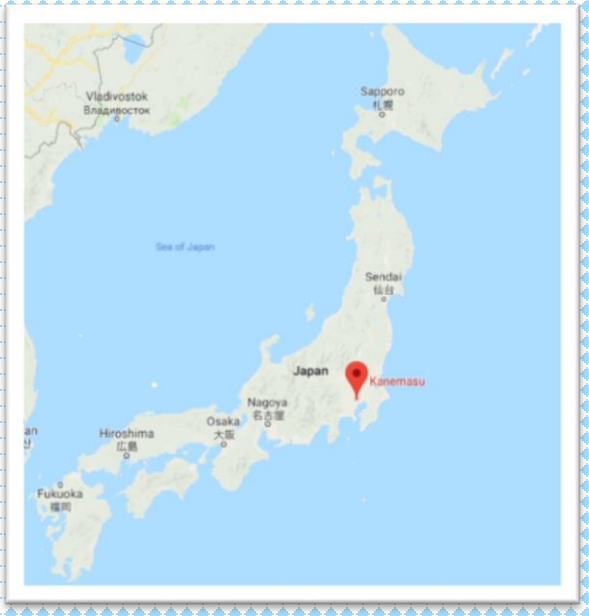
slides[0].shape.text_frame.text
slides[0].shape.left
slides[0].shape.top

text	coordinates
JIST2018 Tutorial	100,2000
Practical use of Knowledge...	100,2350
2018.11.26	100,3970
Fujitsu Laboratories	100,4320

My Favorite Things

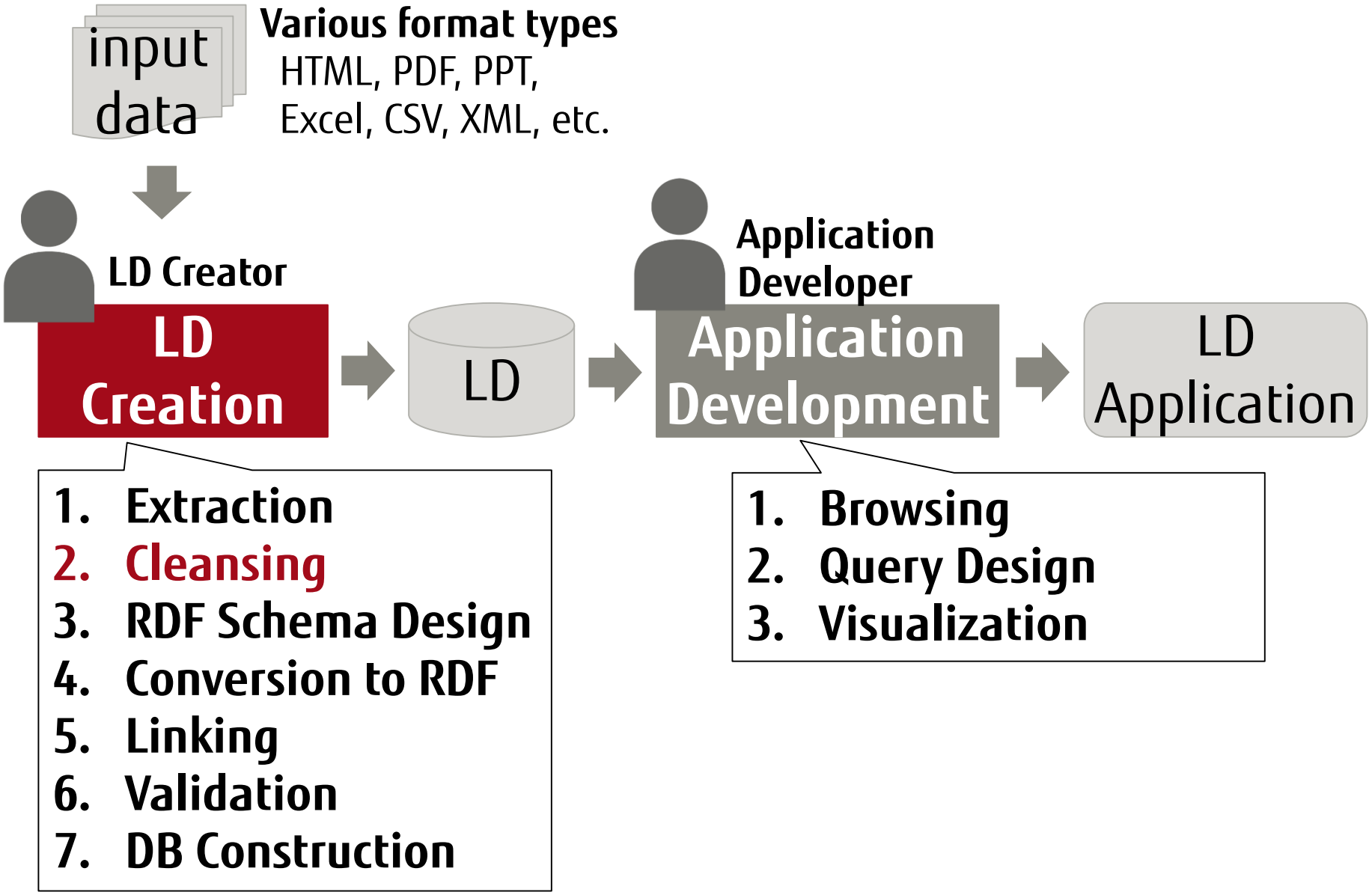
FUJITSU

Sea Urchin and Beef Sashimi Roll



Kanemasu Bar / かねます
Tokyo, Japan
https://tabelog.com/tokyo/A1313/A131302/13002243/

Outline

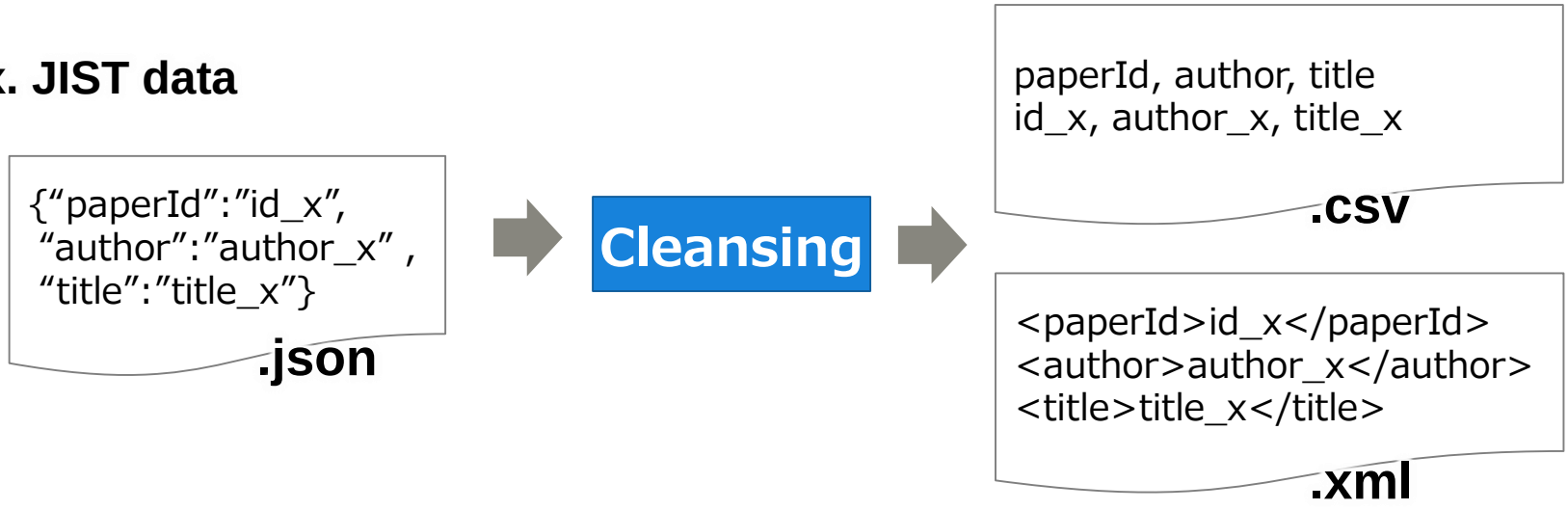


Cleansing

FUJITSU

Cleansing is format conversion to improve the quality and convenience of the extracted data

Ex. JIST data



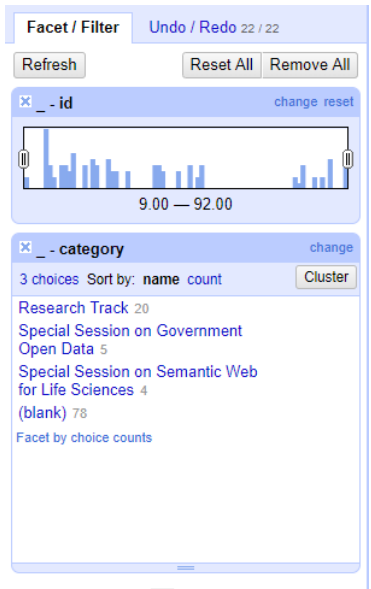
- Cleansing for various data formats
  - OpenRefine★

Cleansing for CSV, XML, JSON

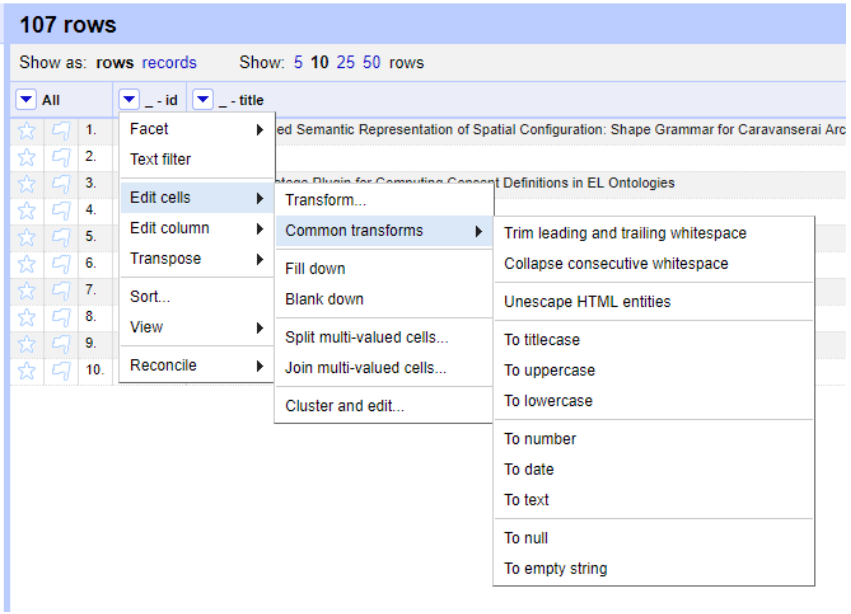
107 rows

Show as: **rows** records Show: 5 **10** 25 50 rows

All		_ - id	_ - title	_ - category	_ - authors - autl
	1.	9	Ontology-based Semantic Representation of Spatial Configuration: Shape Grammar for Caravanserai Architectural Heritage	Research Track	Elham Andarodi
	2.	9			Frederic Andres
	3.	16	DeFind: A Protege Plugin for Computing Concept Definitions in EL Ontologies	Research Track	Denis Ponomaryov
	4.	16			Stepan Yakovenko
	5.	18	Knowledge Driven Intelligent Survey SystemS for Linguists	Research Track	Ricardo Soares
	6.	18			Jeff Z. Pan
	7.	18			Elspeth Edelstein
	8.	18			Adam Wyner
	9.	19	More is Better: Sequential combinations of knowledge graph embedding approaches	Research Track	Kemas Wiharja
	10.	19			Jeff Z. Pan

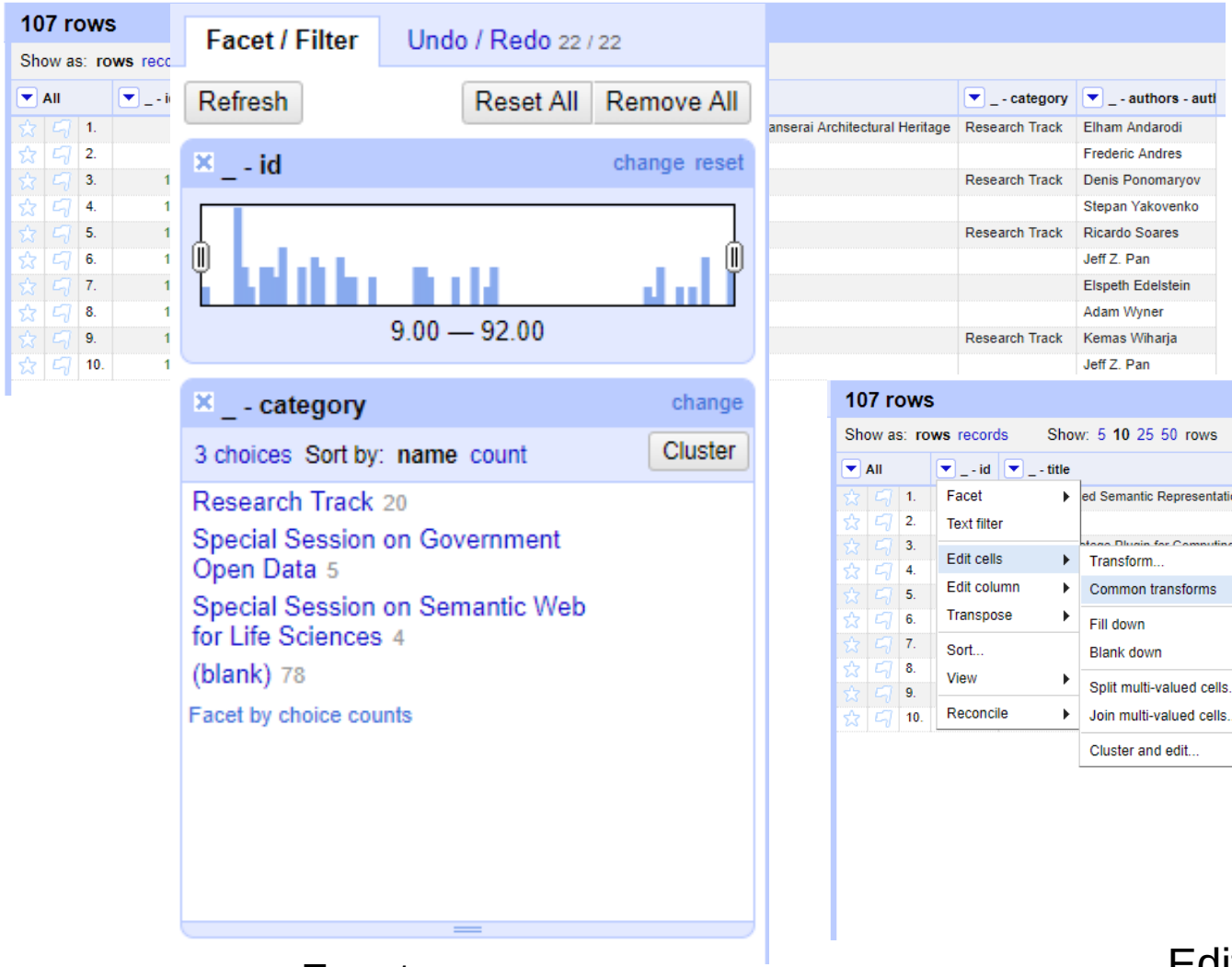


Facet



Edit cells

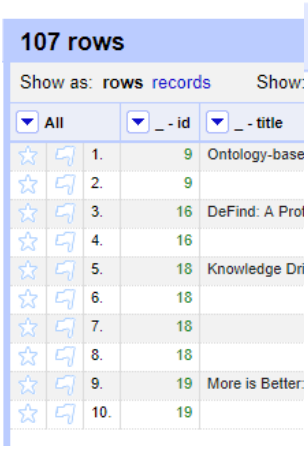
Cleansing for CSV, XML, JSON

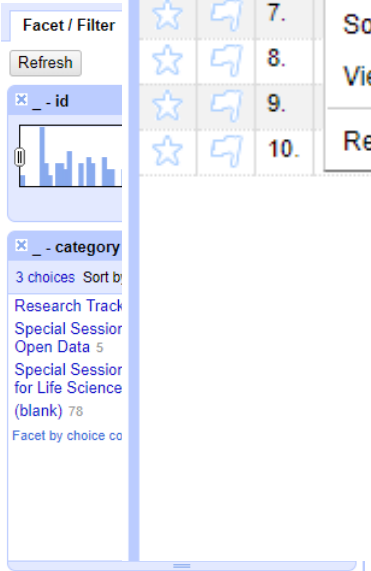


Facet

Edit cells

Cleansing for CSV, XML, JSON





Facet

Edit cells

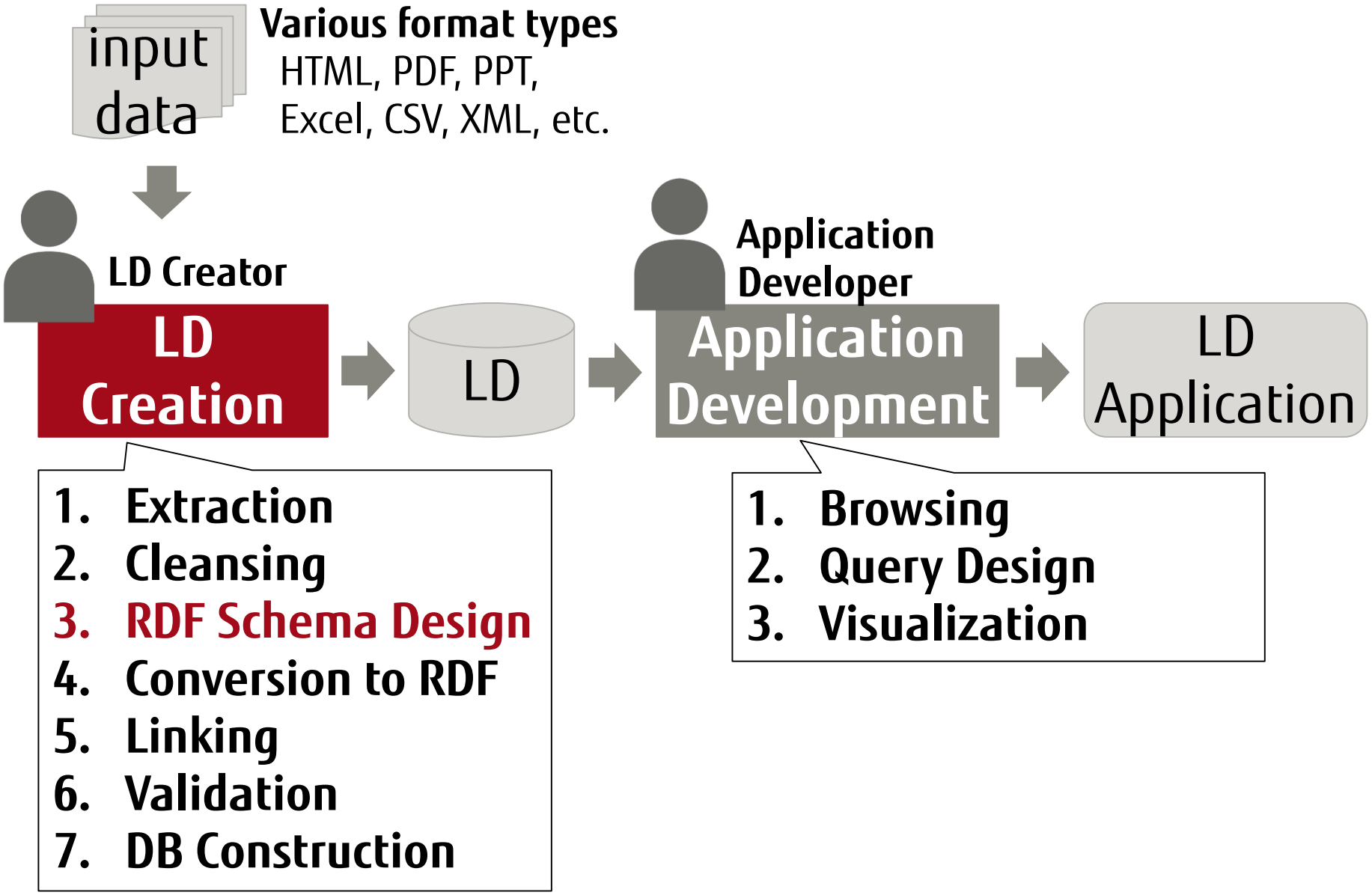
My Favorite Things

FUJITSU

Bungee Jump

Minakami Camping Area, Japan
/ みなかみキャンプ場

Outline



RDF Schema Design

FUJITSU

Design the schema for RDF conversion of the extracted / cleansed data, and construct an ontology as necessary

Ex. JIST data

- Searching
  - Vocabulary / Ontology: Linked Open Vocabularies★
  - Dataset: Google Dataset Search★
  - Instance: LOD Laundromat, LOD4ALL
- Browsing
  - Data Sniffer★
- Building
  - Protégé

Linked Open Vocabularies ★

https://lov.linkeddata.es/

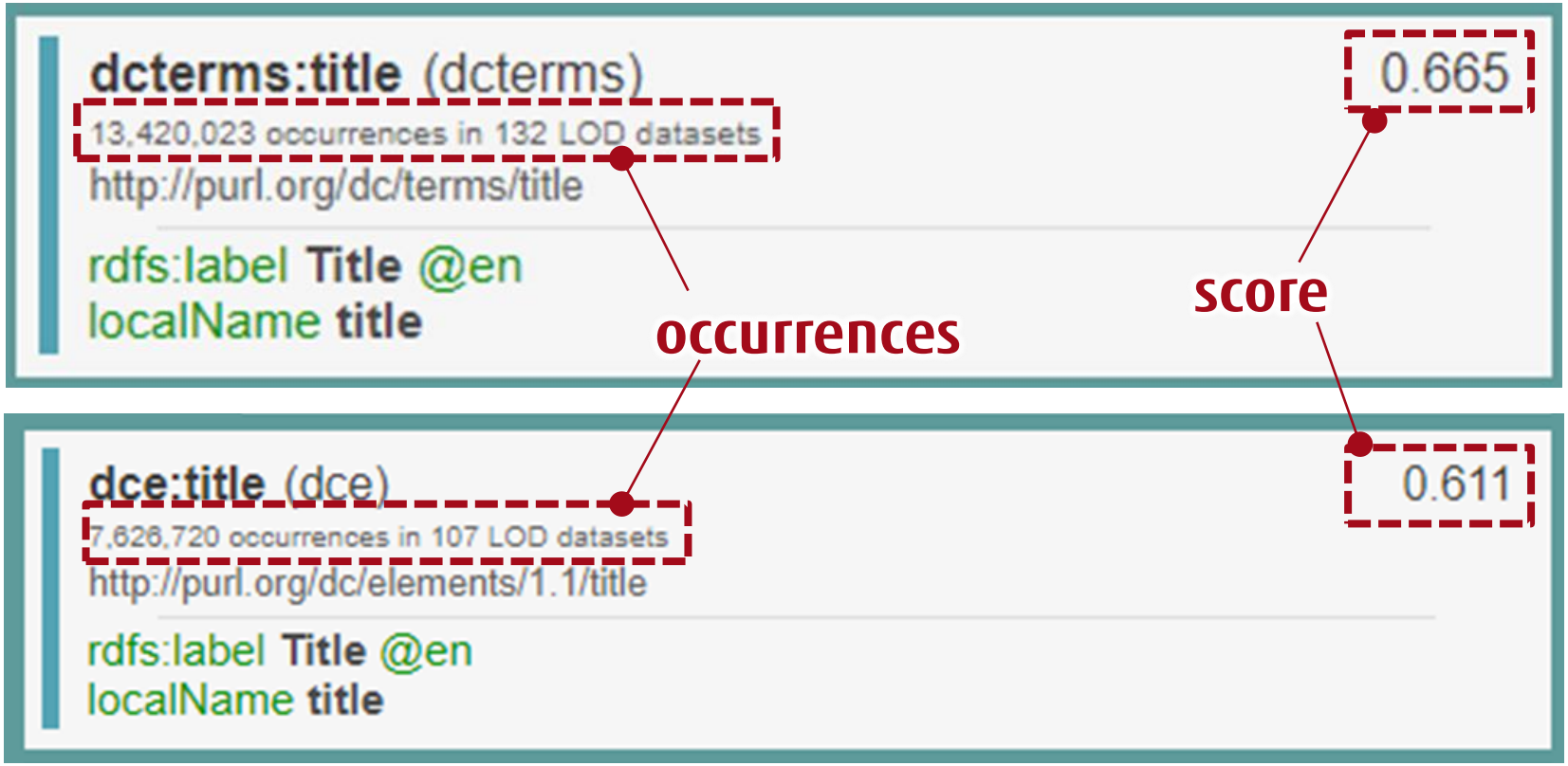
FUJITSU

- A service to search for existing vocabulary
  - 651 vocabularies are included (*October 2018)

Example: ”research article title”

Linked Open Vocabularies ★

https://lov.linkeddata.es/



dcterms:title / dce:title are most widely used

*dcterms:title is a sub property of dce:title.

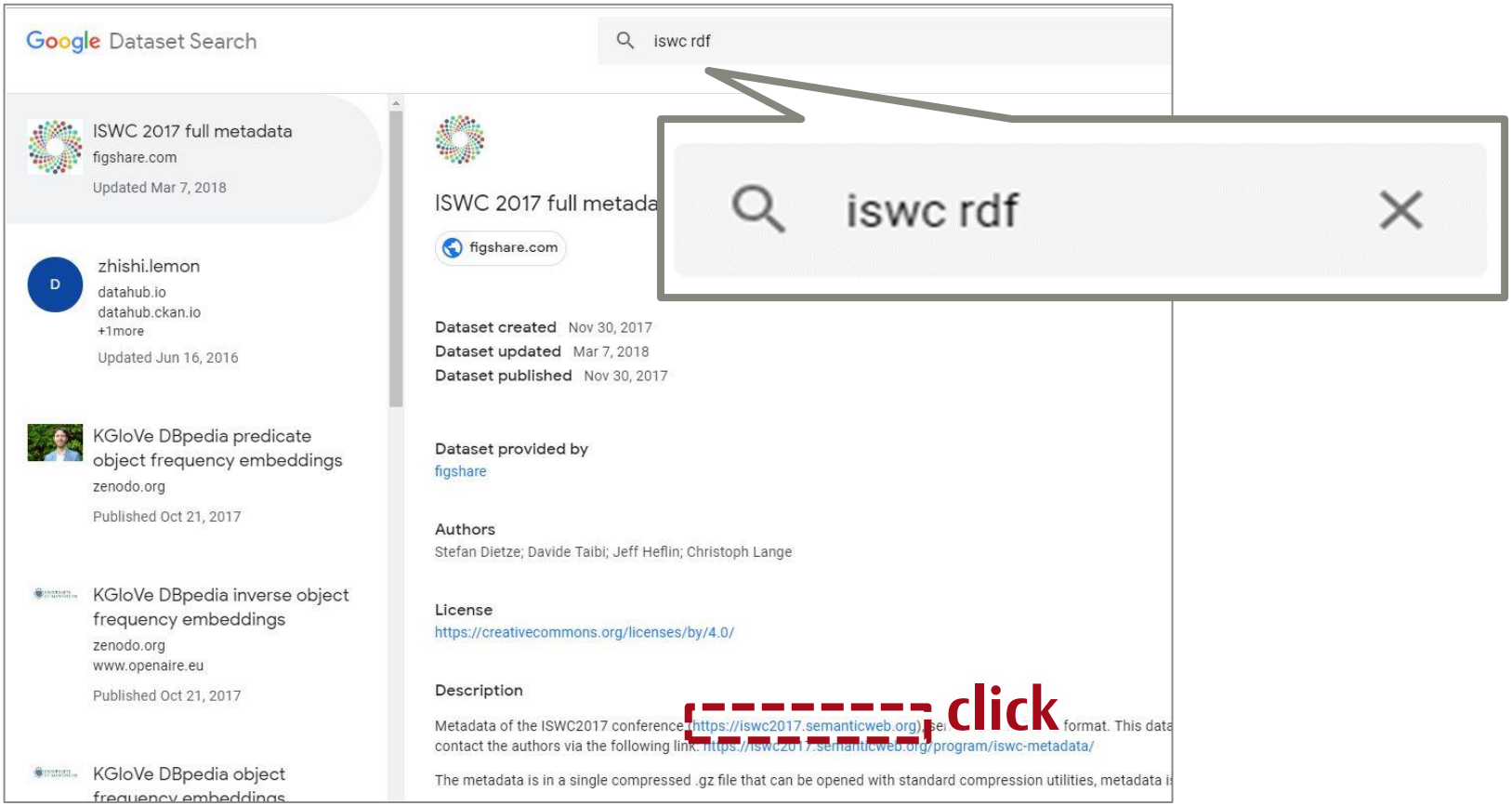
Google Dataset Search ★

https://toolbox.google.com/datasetsearch

FUJITSU

- Dataset search service by Google (launched September, 2018)
  - Enables searching of datasets published on the web.

Example: ISWC RDF data



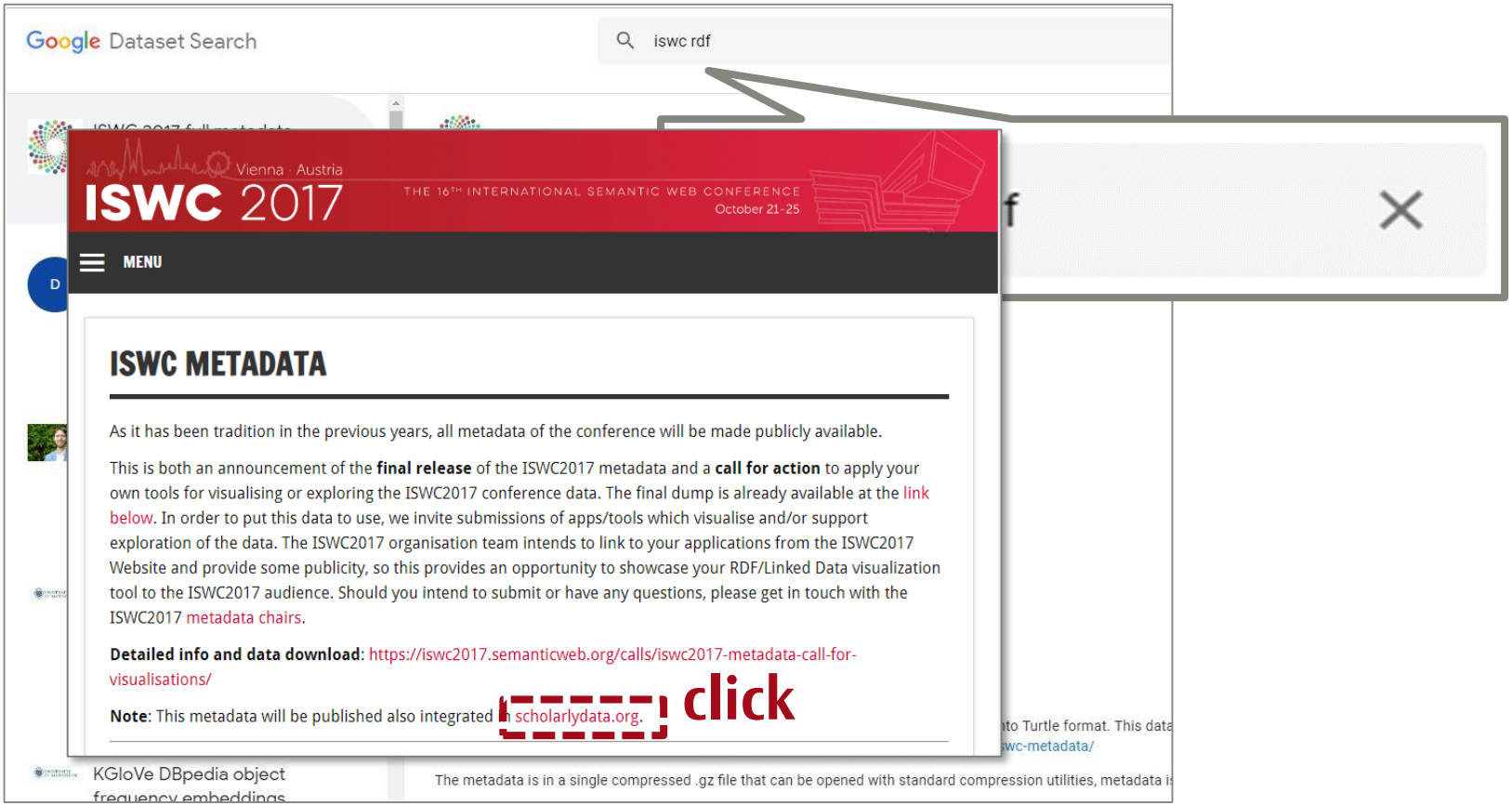
Google Dataset Search ★

https://toolbox.google.com/datasetsearch

FUJITSU

- Dataset search service by Google (launched September, 2018)
  - Enables searching of datasets published on the web.

Example: ISWC RDF data



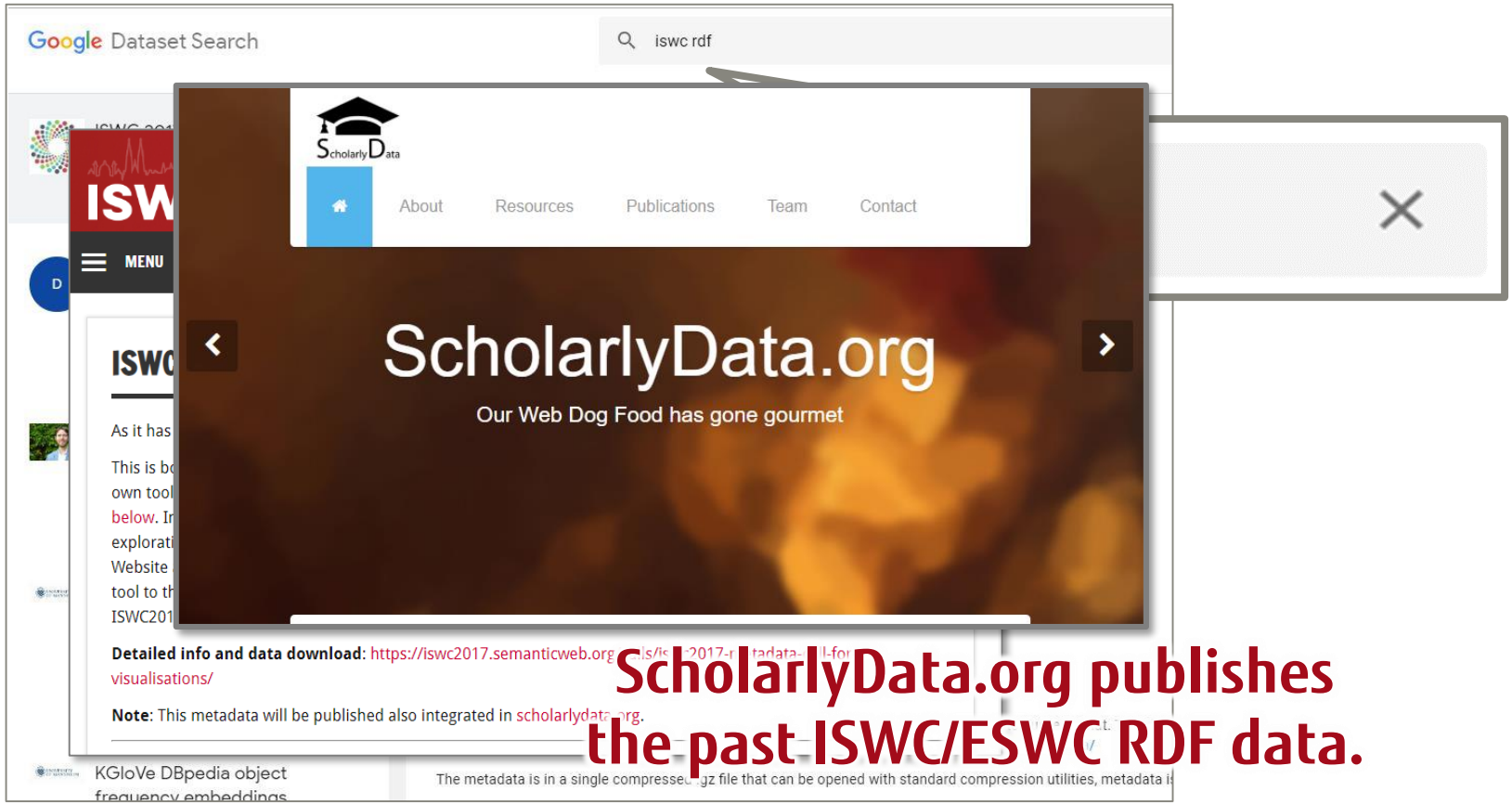
Google Dataset Search ★

https://toolbox.google.com/datasetsearch

FUJITSU

- Dataset search service by Google (launched September, 2018)
  - Enables searching of datasets published on the web.

Example: ISWC RDF data

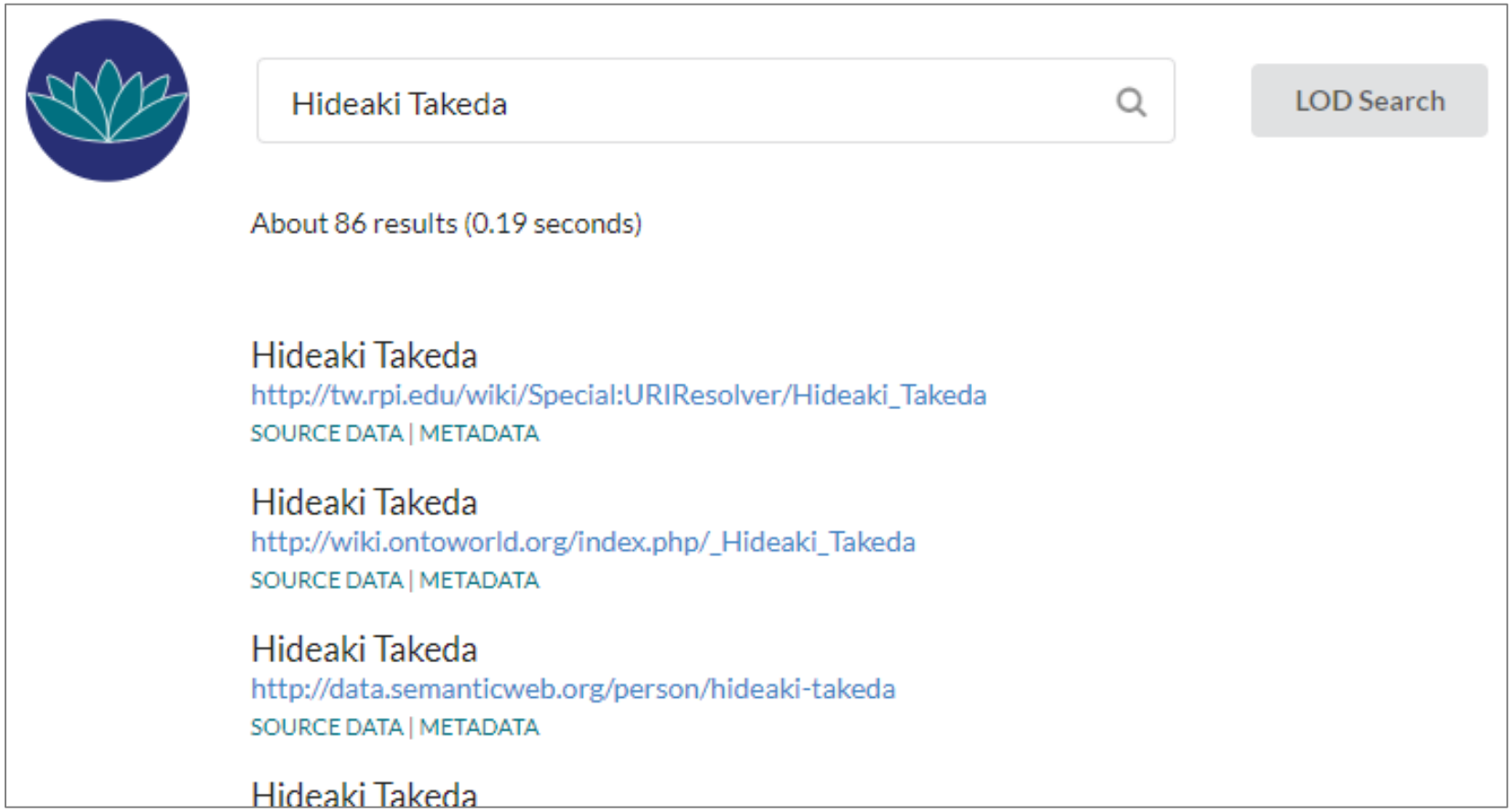


ScholarlyData.org publishes the past ISWC/ESWC RDF data.

LOD Laundromat

http://lodlaundromat.org/

- Crawling from the LOD Cloud and converting it to beautiful data
  - Removes duplicates, syntax errors, and empty nodes
- SPARQL endpoint and entity search function (LOTUS) also provided



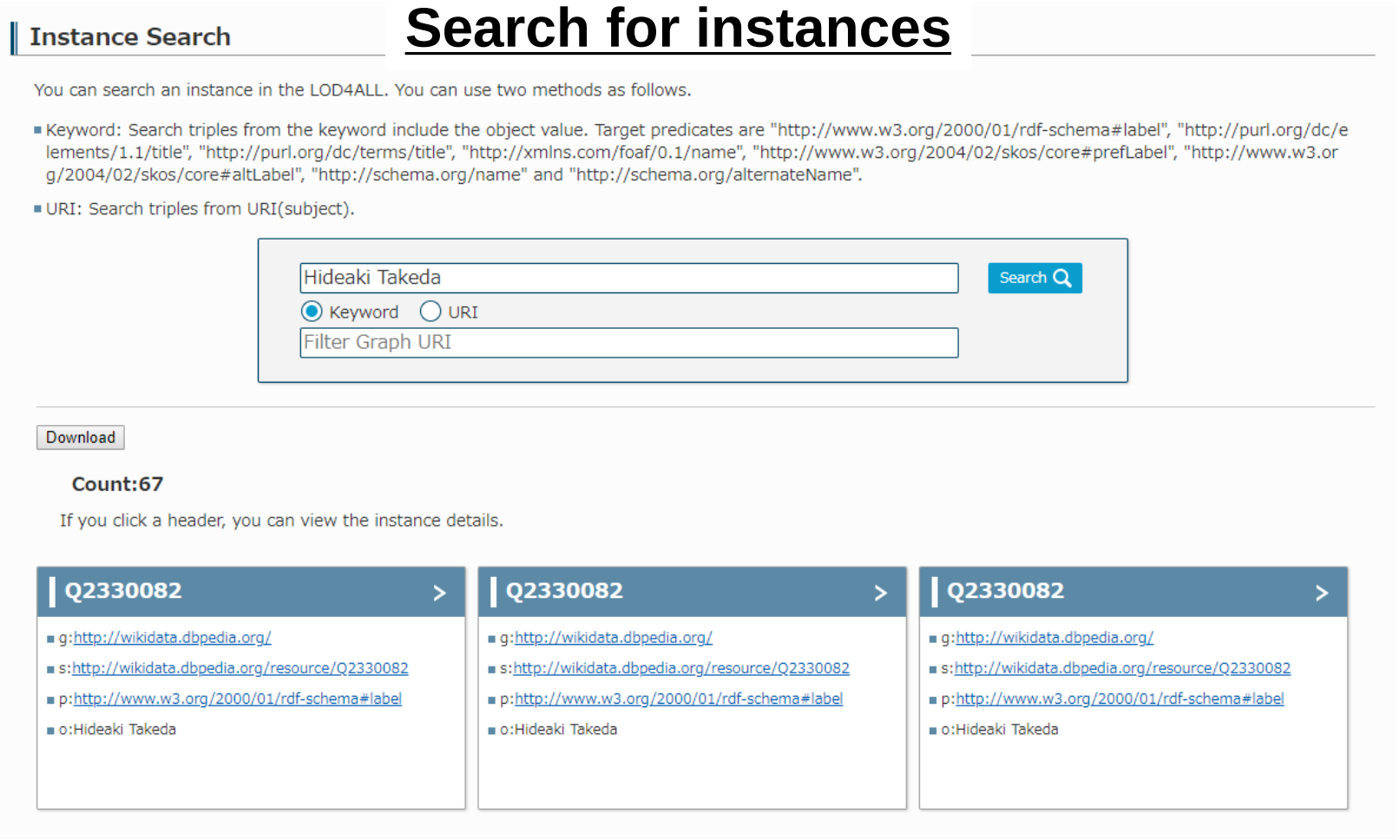
LOD4ALL

RDF Schema Design(4/6)

https://lod4all.net/index.html

FUJITSU

- A service delivering a single-stop entry point for LOD utilization
  - Browsing, searching, and accessing capability for publicly available Linked Data.
- LOD4ALL also provides a development platform for applications using LOD



OpenLink Structured Data Sniffer ★

FUJITSU

- Enable to confirm RDF and metadata in HTML
  - Chrome extension

DBLP website

[–] Publication search results

found 36 matches

2017

Franz Baader, Daniel Borchmann, Adrian Nuradiansyah:
dentity Problem in Description Logic Ontologies and Its Application to View-Based Information
g. JIST 2017: 102-117

export record
BibTeX
RDF N-Triples
XML
dblp key:
conf/jist/BaaderBN17

Turtle

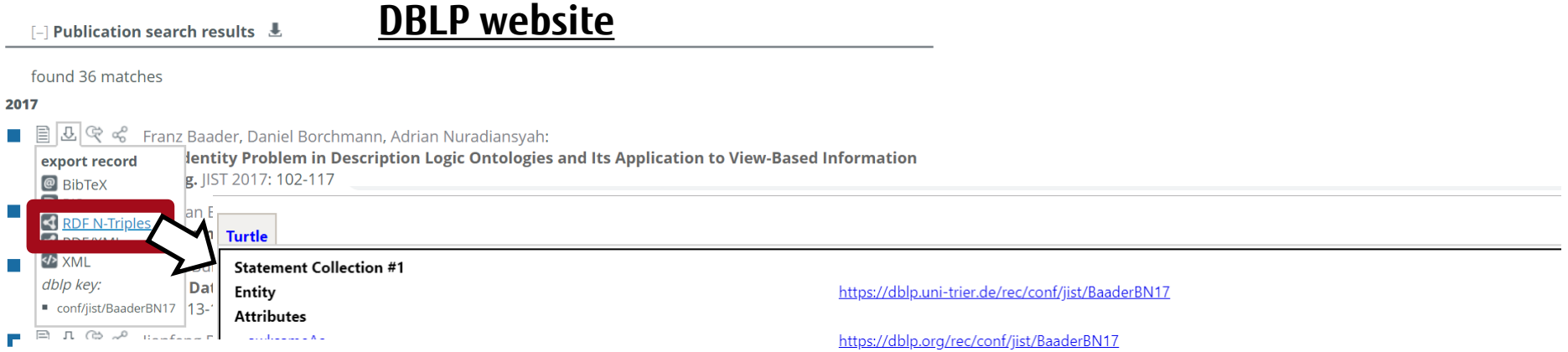
Statement Collection #1	
Entity	https://dblp.uni-trier.de/rec/conf/jist/BaaderBN17
Attributes	
owl:sameAs	https://dblp.org/rec/conf/jist/BaaderBN17
Statement Collection #2	
Entity	https://dblp.org/rec/conf/jist/BaaderBN17
Attributes	
rdf:type	https://dblp.uni-trier.de/rdf/schema-2017-04-18#Publication
owl:sameAs	https://doi.org/10.1007/978-3-319-70682-5_7
https://dblp.uni-trier.de/rdf/schema-2017-04-18#title	The Identity Problem in Description Logic Ontologies and Its Application to View-Based Information Hiding.
https://dblp.uni-trier.de/rdf/schema-2017-04-18#bibtexType	http://data.bibbase.org/ontology/#Inproceedings
https://dblp.uni-trier.de/rdf/schema-2017-04-18#publicationType	https://dblp.uni-trier.de/rdf/schema-2017-04-18#Inproceedings
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/b/Baader:Franz
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/b/Borchmann:Daniel
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/n/Nuradiansyah:Adrian
https://dblp.uni-trier.de/rdf/schema-2017-04-18#primaryElectronicEdition	https://doi.org/10.1007/978-3-319-70682-5_7
https://dblp.uni-trier.de/rdf/schema-2017-04-18#listedOnTocPage	https://dblp.org/db/conf/jist/jist2017
https://dblp.uni-trier.de/rdf/schema-2017-04-18#publishedInBook	JIST
https://dblp.uni-trier.de/rdf/schema-2017-04-18#pageNumbers	102-117
https://dblp.uni-trier.de/rdf/schema-2017-04-18#yearOfPublication	2017
https://dblp.uni-trier.de/rdf/schema-2017-04-18#publishedAsPartOf	https://dblp.org/rec/conf/jist/2017
Statement Collection #3	
Entity	https://dblp.org/rec/conf/jist/BaaderBN17.nt

OpenLink Structured Data Sniffer

FUJITSU

- Chrome extension developed by OpenLink
  - Enable to confirm RDF and metadata in HTML

DBLP website



Statements of one publication of JIST 2017:

Entity	https://dblp.org/rec/conf/jist/BaaderBN17
Attributes	
rdf:type	https://dblp.uni-trier.de/rdf/schema-2017-04-18#Publication
owl:sameAs	https://doi.org/10.1007/978-3-319-70682-5_7
https://dblp.uni-trier.de/rdf/schema-2017-04-18#title	The Identity Problem in Description Logic Ontologies and Its A
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/b/Baader:Franz
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/b/Borchmann:Daniel
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/n/Nuradiansyah:Adrian
Entity	https://dblp.org/rec/conf/jist/BaaderBN17.nt

DBLP's own vocabulary

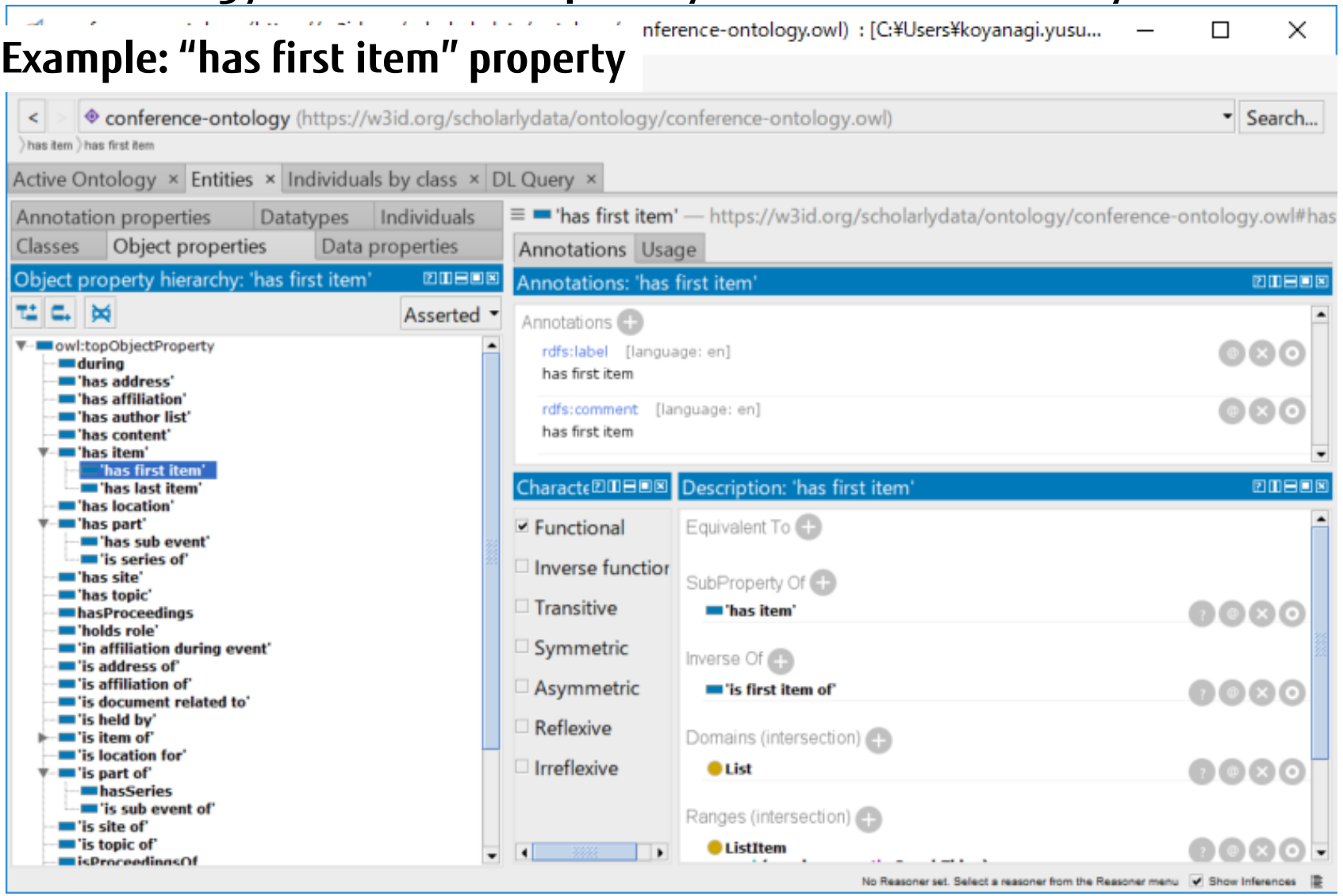
Protégé

https://protege.stanford.edu/

FUJITSU

- An ontology editor developed by Stanford University

Example: “has first item” property

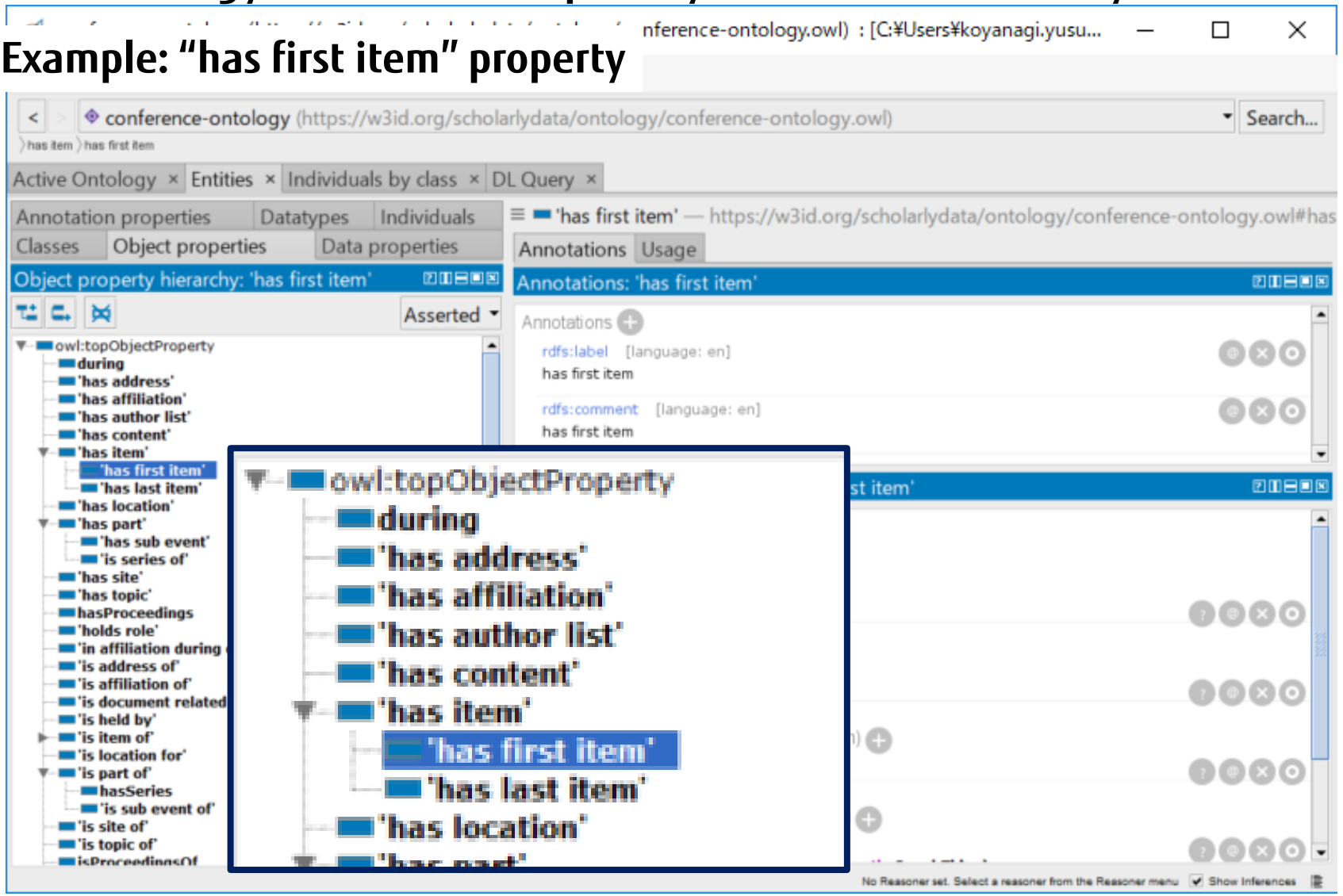


Protégé

https://protege.stanford.edu/

- An ontology editor developed by Stanford University

Example: "has first item" property

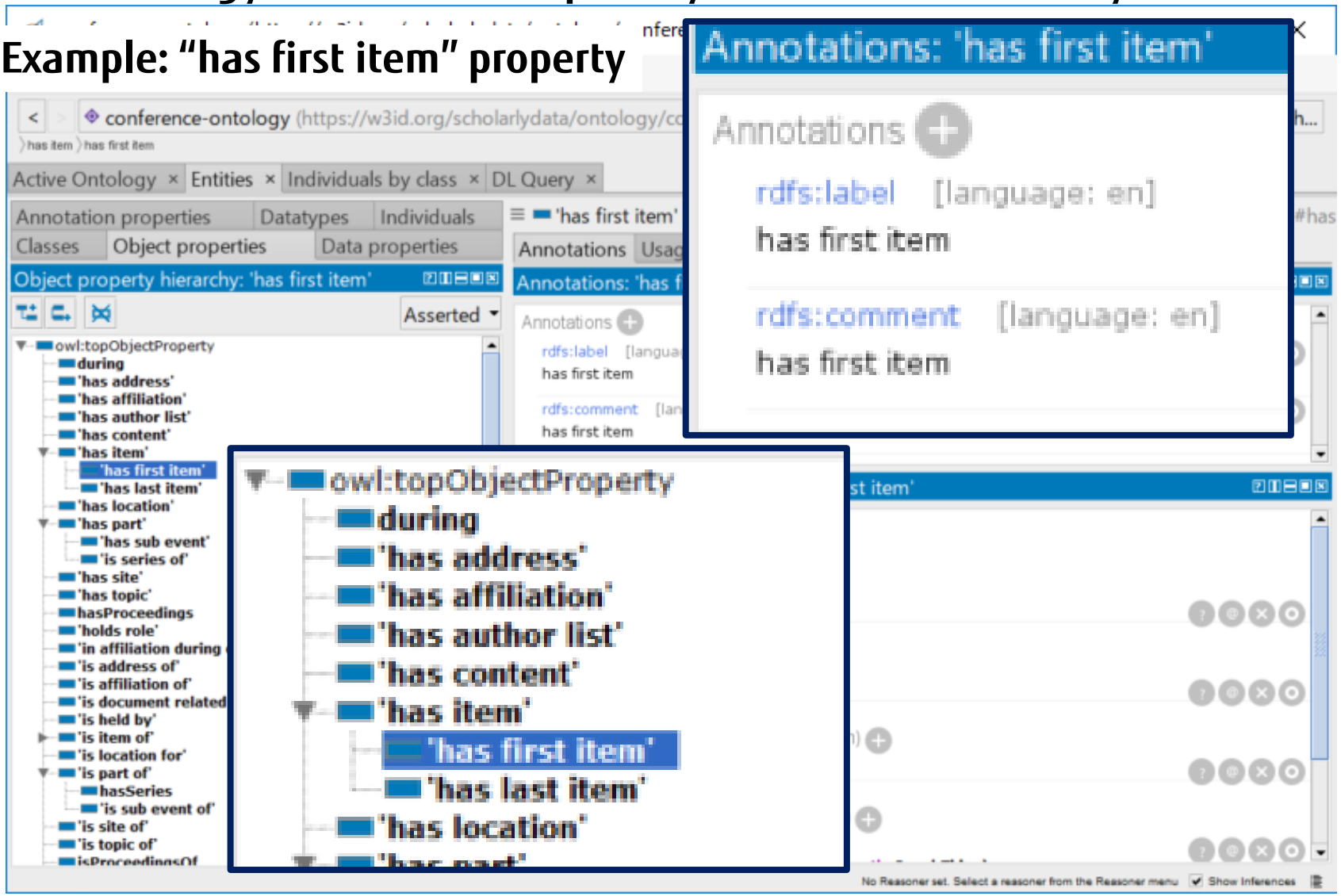


Protégé

https://protege.stanford.edu/

- An ontology editor developed by Stanford University

Example: “has first item” property

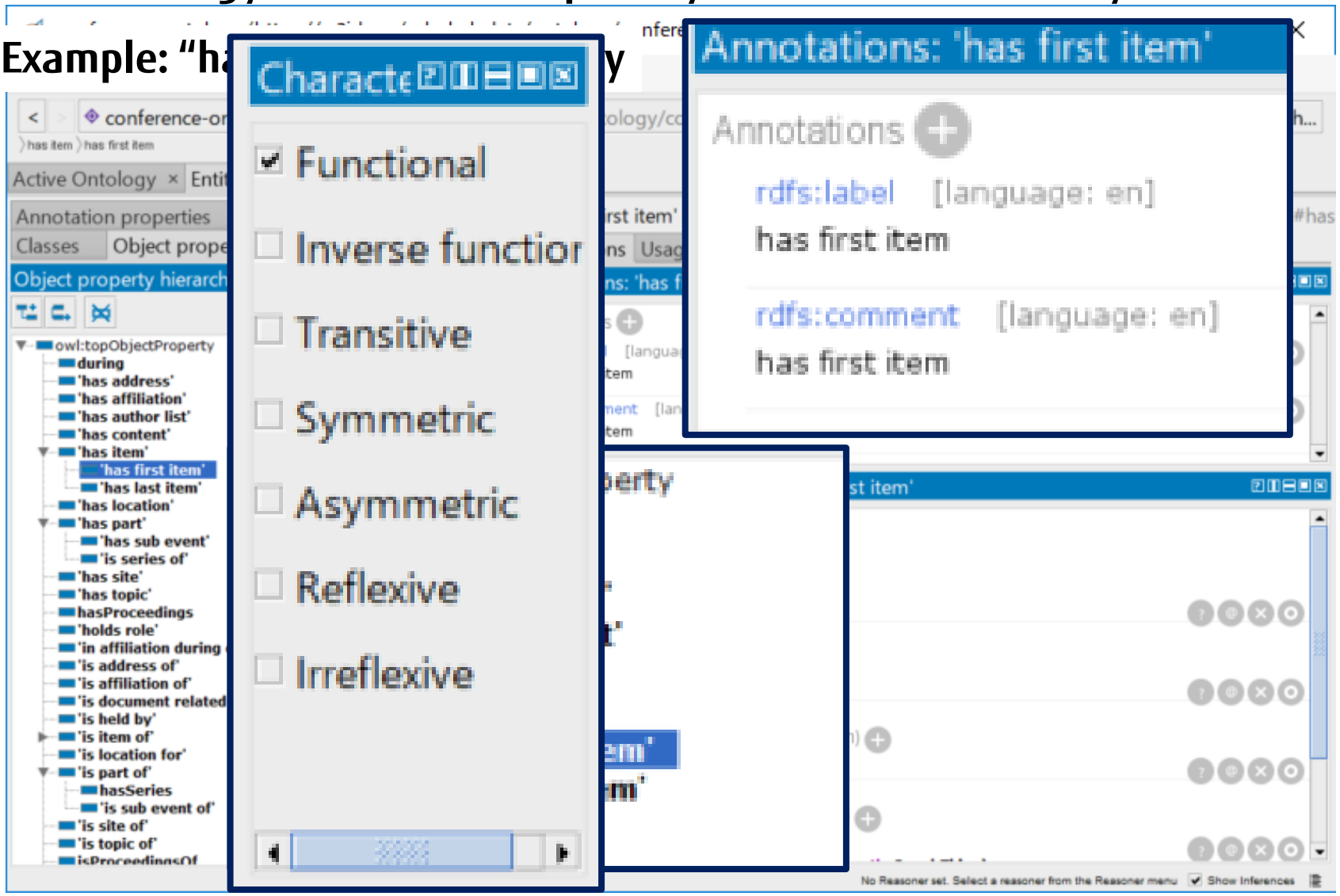


Protégé

https://protege.stanford.edu/

- An ontology editor developed by Stanford University

Example: "ha



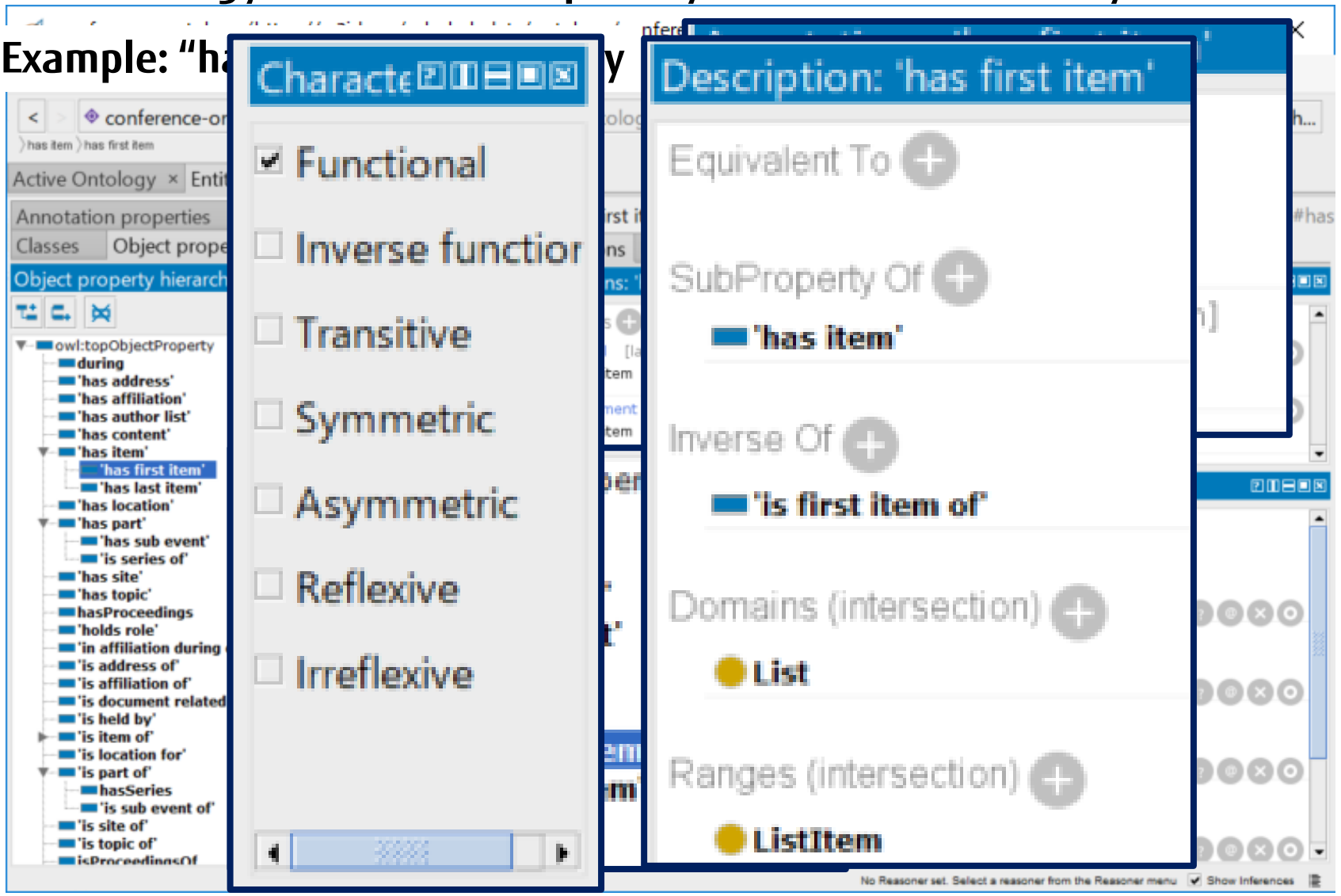
Protégé

https://protege.stanford.edu/

FUJITSU

- An ontology editor developed by Stanford University

Example: "ha

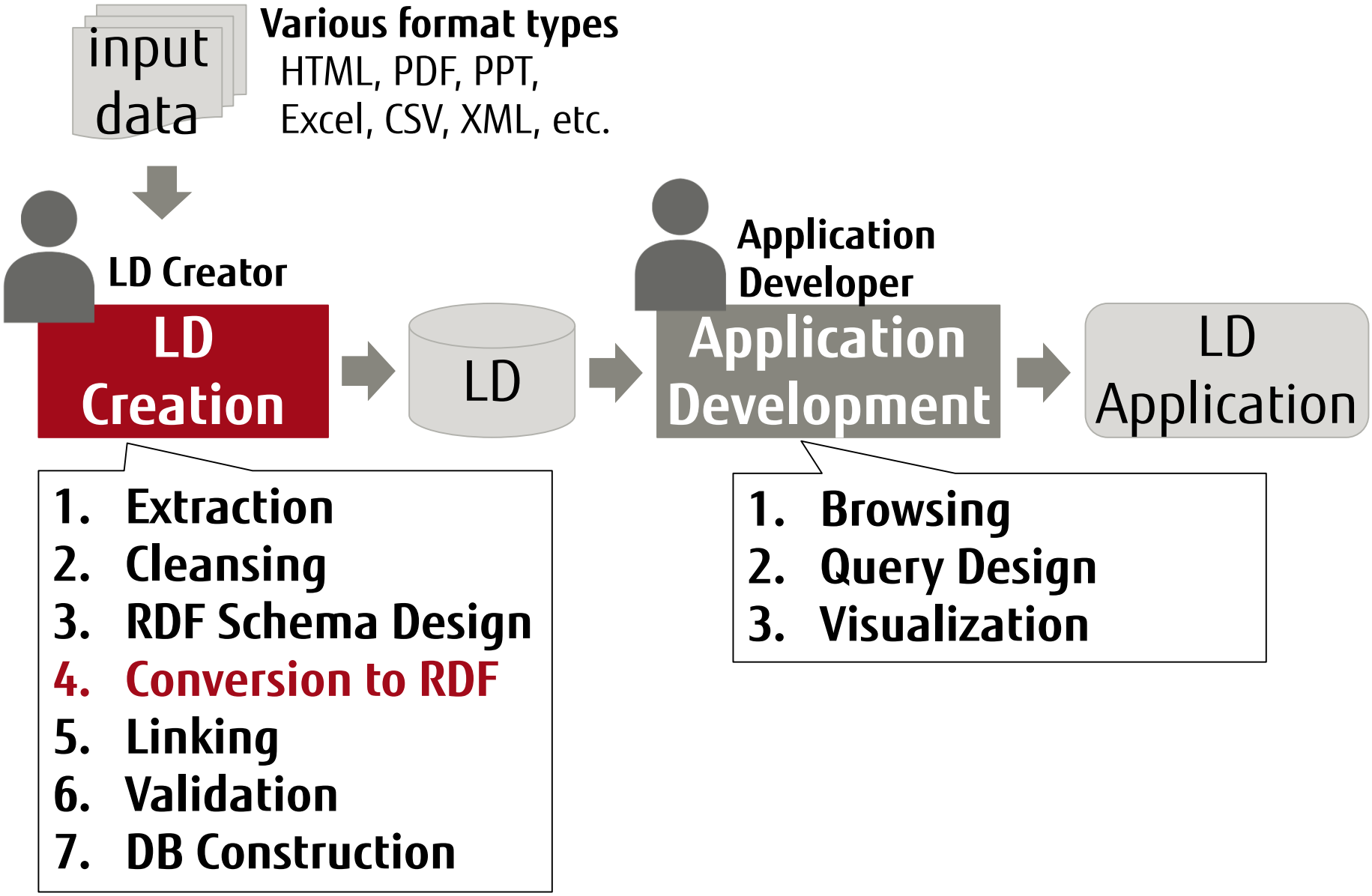


My Favorite Things

FUJITSU

Oilfish / Escolar Sashimi

Outline

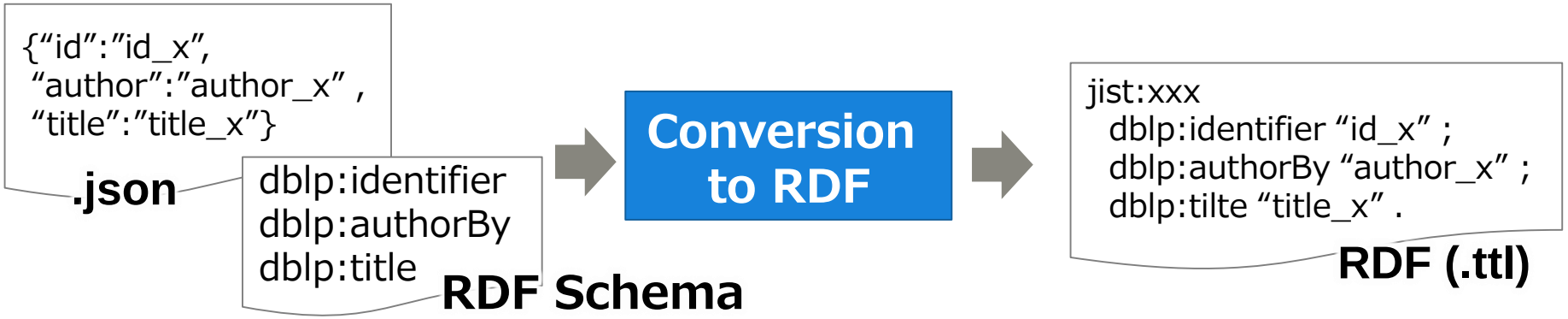


Conversion to RDF

FUJITSU

Convert an original dataset and an RDF schema to RDF format

Ex. JIST data



- JSON/CSV/XML to RDF
  - Open Refine + RDF Extension★
  - YARRRML
- CSV to RDF
  - tarql

Open Refine + RDF Extension

Conversion to RDF (1/3)

https://github.com/sparkica/rdf-extension

- The skeleton specifies how the RDF data will be generated from your grid-shaped data.

Open Refine + RDF Extension

Conversion to RDF (1/3)

FUJITSU

https://github.com/sparkica/rdf-extension

- Human-readable text-based representation for declarative Linked Data generation rules.
  - A subset of YAML, a widely used data serialization language designed to be human-friendly.
  - Can already be used to represent R2RML and RML rules.

```
prefixes:
jist: https://lod4all.github.io/jist2018-tutorial/jist.ttl#
dblp: https://dblp.org/rdf/schema-2017-04-18#
mappings:
 accepted_papers:
   sources:
     - [‘jist.json~jsonpath', '$.accepted_papers[*]']
   s: https://lod4all.github.io/jist2018-tutorial/jist.ttl#$(id)
   po:
     - [a, dblp:Publication]
     - [dblp:authoredBy, jist:$(authors)~iri]
     - [dblp:title, $(title)]
```

Conversion to RDF (3/3)

https://github.com/tarql/tarql

FUJITSU

- Command-line tool for converting CSV files to RDF
  - Uses SPARQL 1.1 syntax
  - Written in Java and based on Apache ARQ.

- Usage

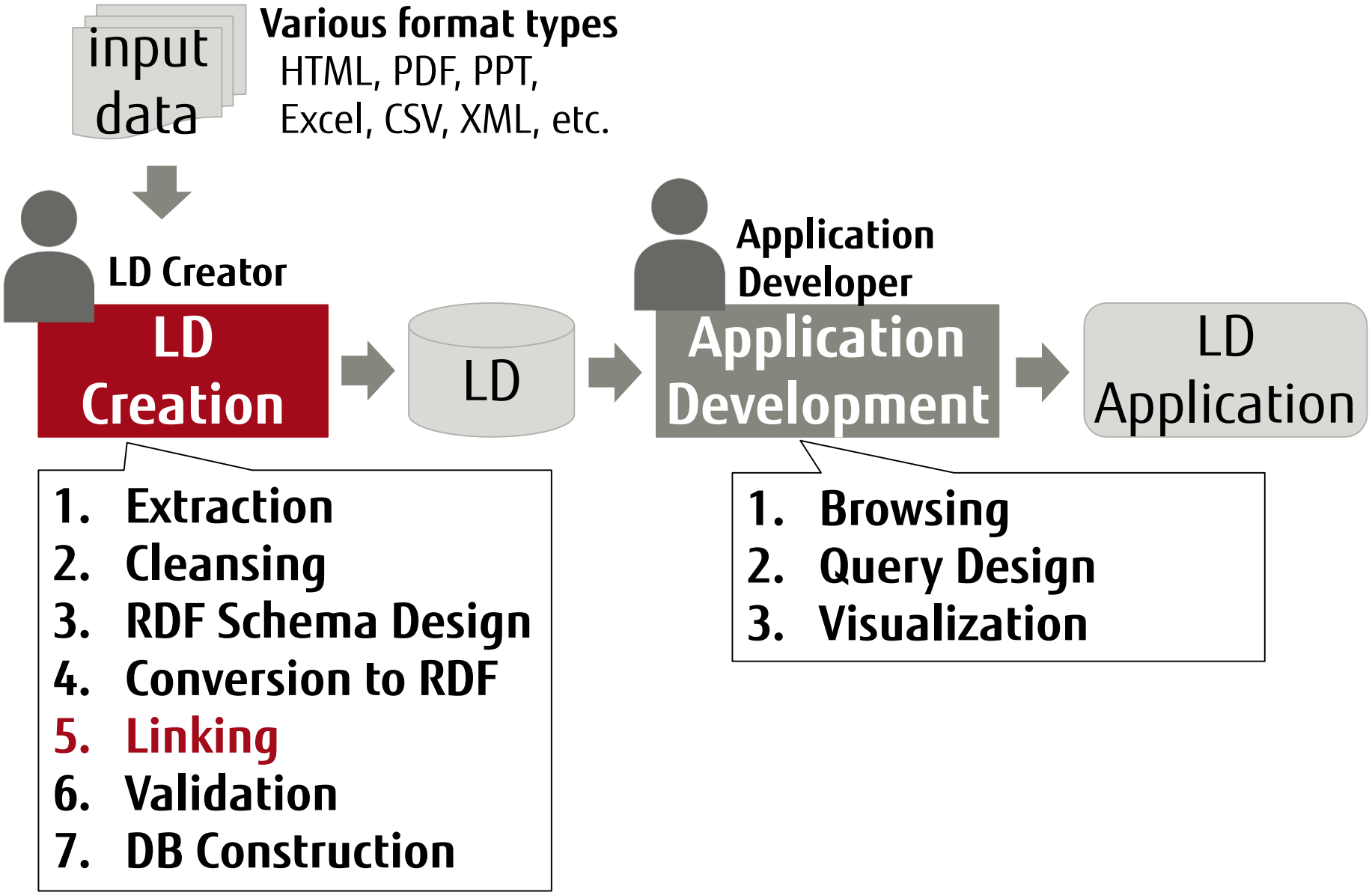
```
CONSTRUCT {
    ?URI a dblp:Publication;
        dblp:authoredBy ?AUTHOR;
        dc:title ?title ;
        dc:category ?category ;
  }
  FROM <file:jist.csv>
  WHERE {
    BIND (URI(?id) AS ?URI)
    BIND (URI(?authors)
AS ?AUTHOR)
}
```

My Favorite Things

FUJITSU

Sky Diving(Hawaii)

Outline

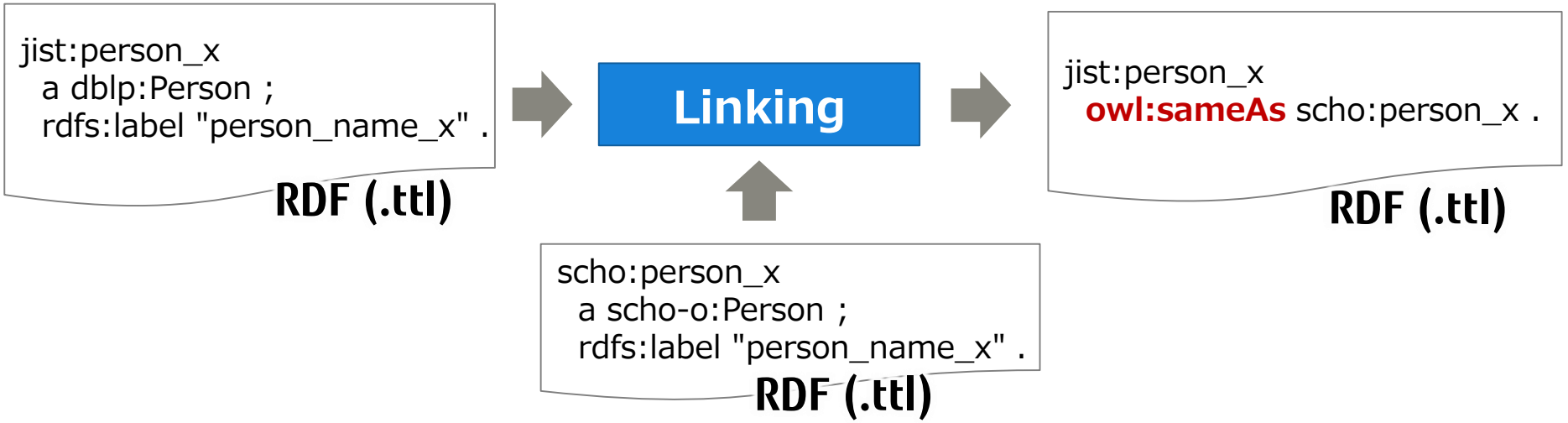


Linking

FUJITSU

Link RDF data with another dataset
in order to enable cross-sectional search by SPARQL

Ex. JIST data



- Liking tool
  - Silk★

- Integration tool for heterogeneous data sources

Workspace: UI to make and set up projects

JIST rdf dump

ISWC2017 rdf dump

Dataset for linking result (no triple, at first)

Linking task

You can register datasets such as RDF dumps and SPARQL endpoints.

Integration tool for heterogeneous data sources

Workspace: UI to make and set up projects

Linking task

You can register datasets such as RDF dumps and SPARQL endpoints.

Editor: UI for defining the flow of the Linking task

Source

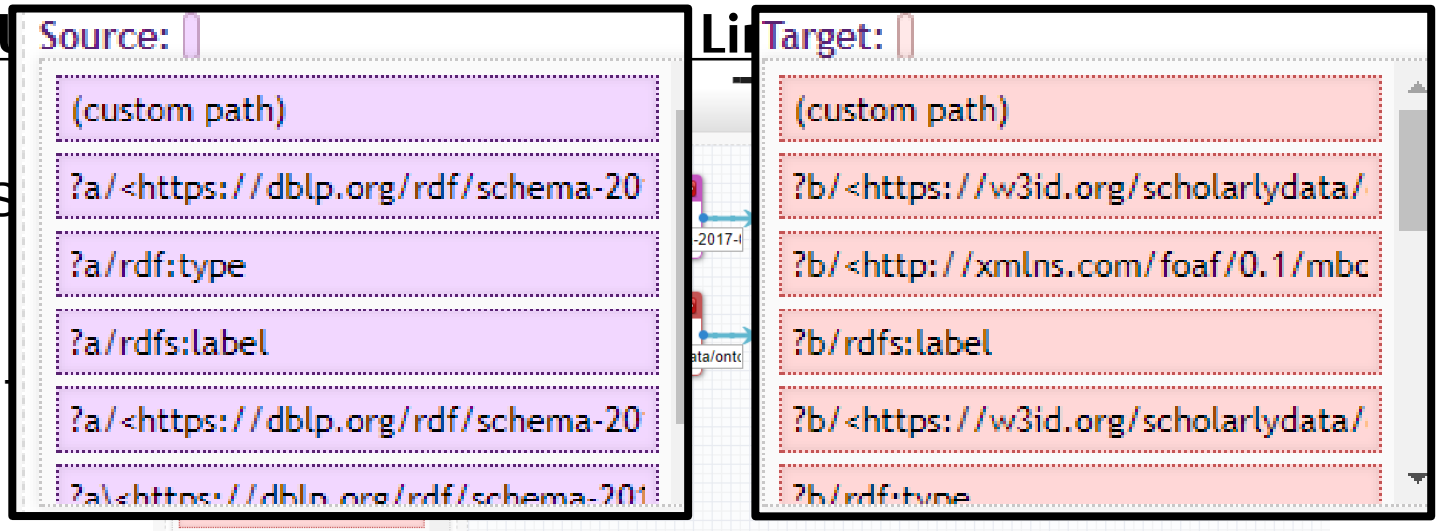
Target

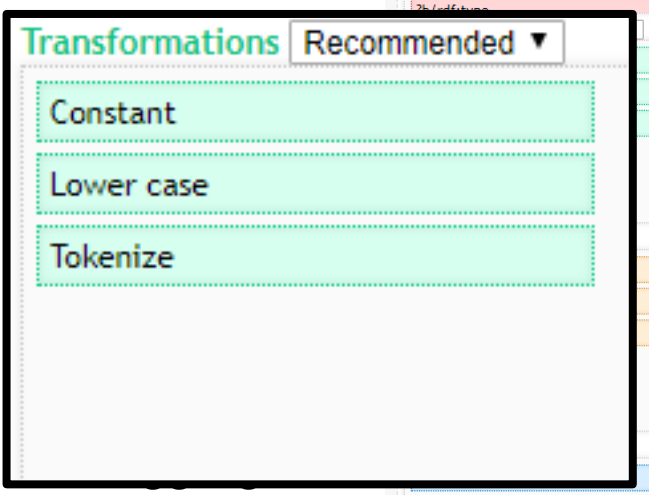
Transformations

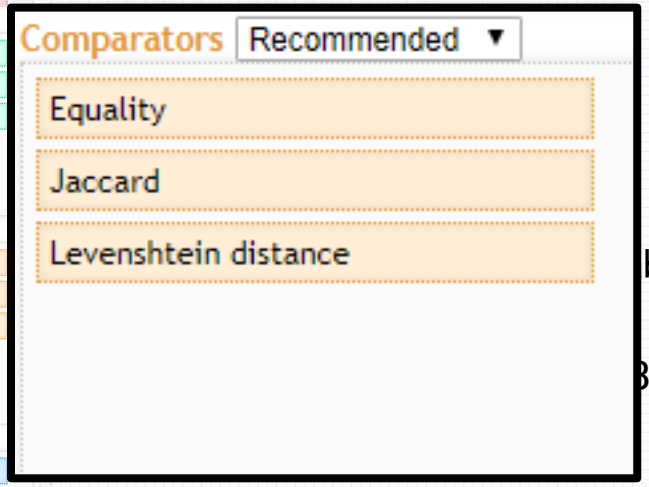
Comparators

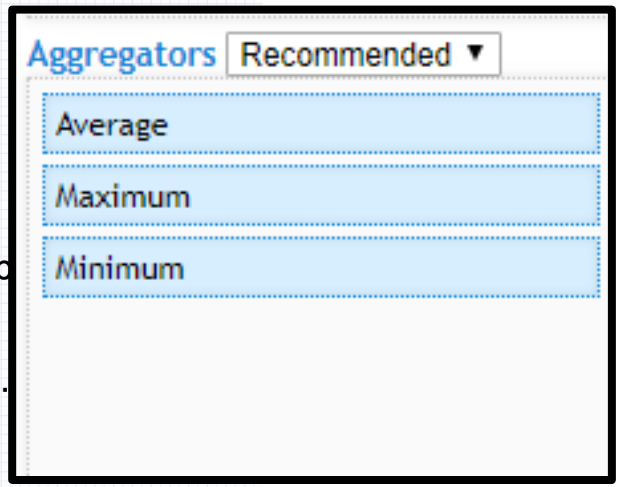
Aggregators

Editor: L





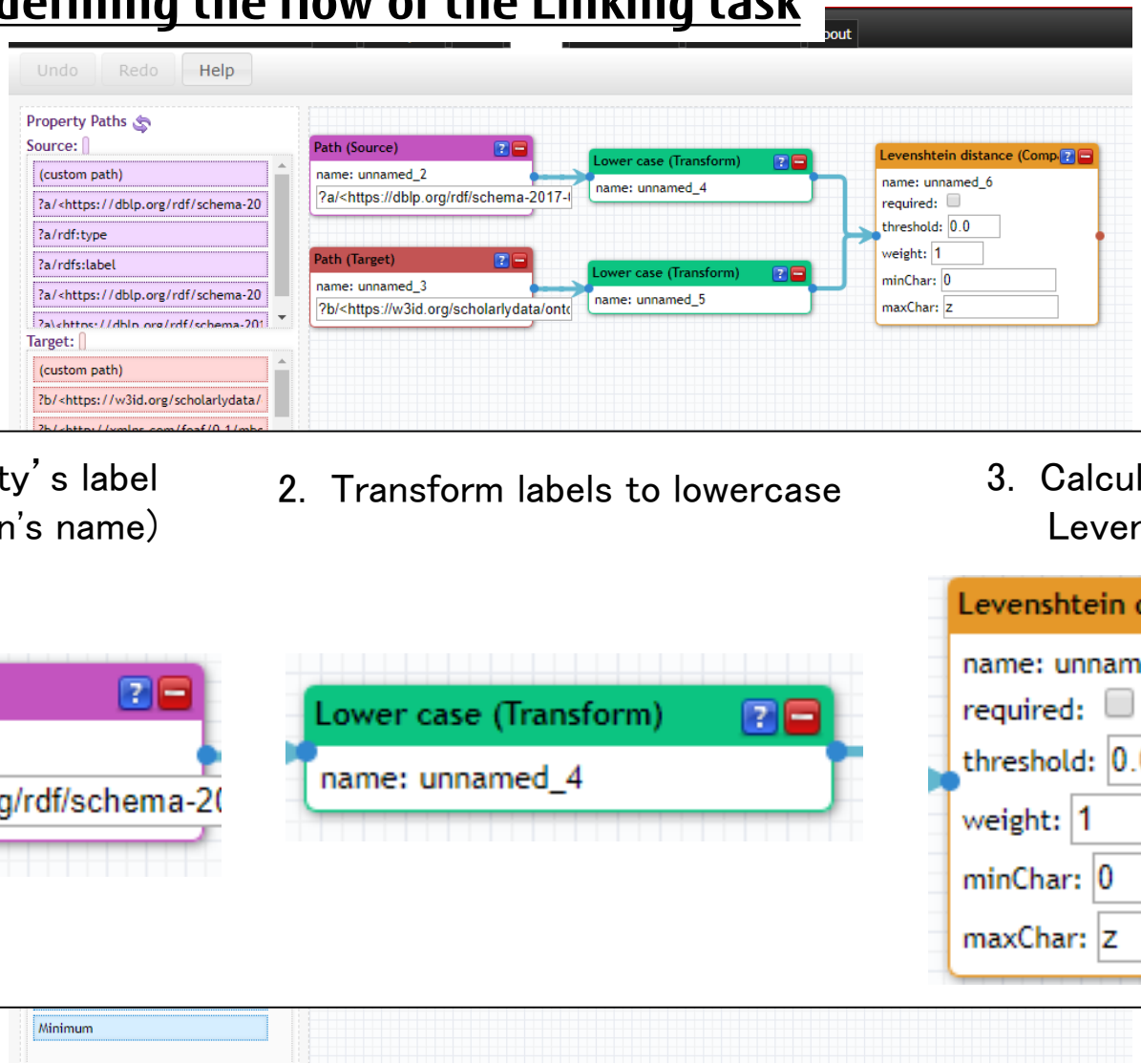




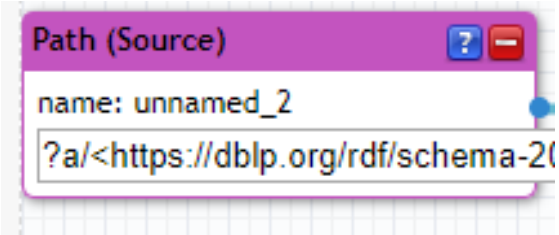
Editor: UI for defining the flow of the Linking task

Source

Target

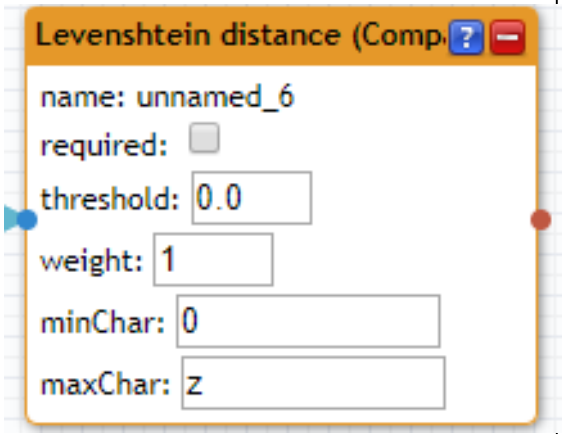


1. Retrieve entity's label (person's name)

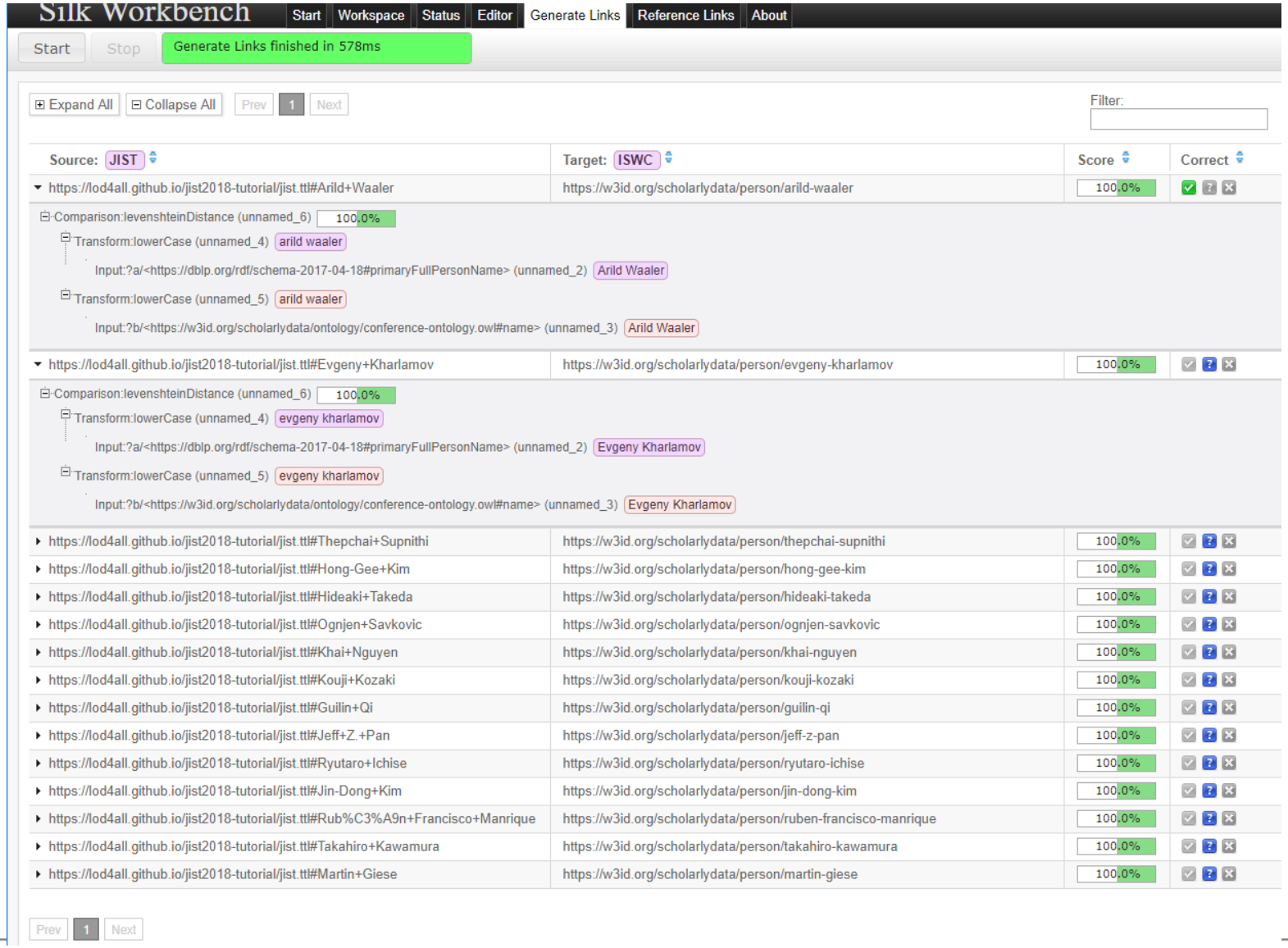


2. Transform labels to lowercase

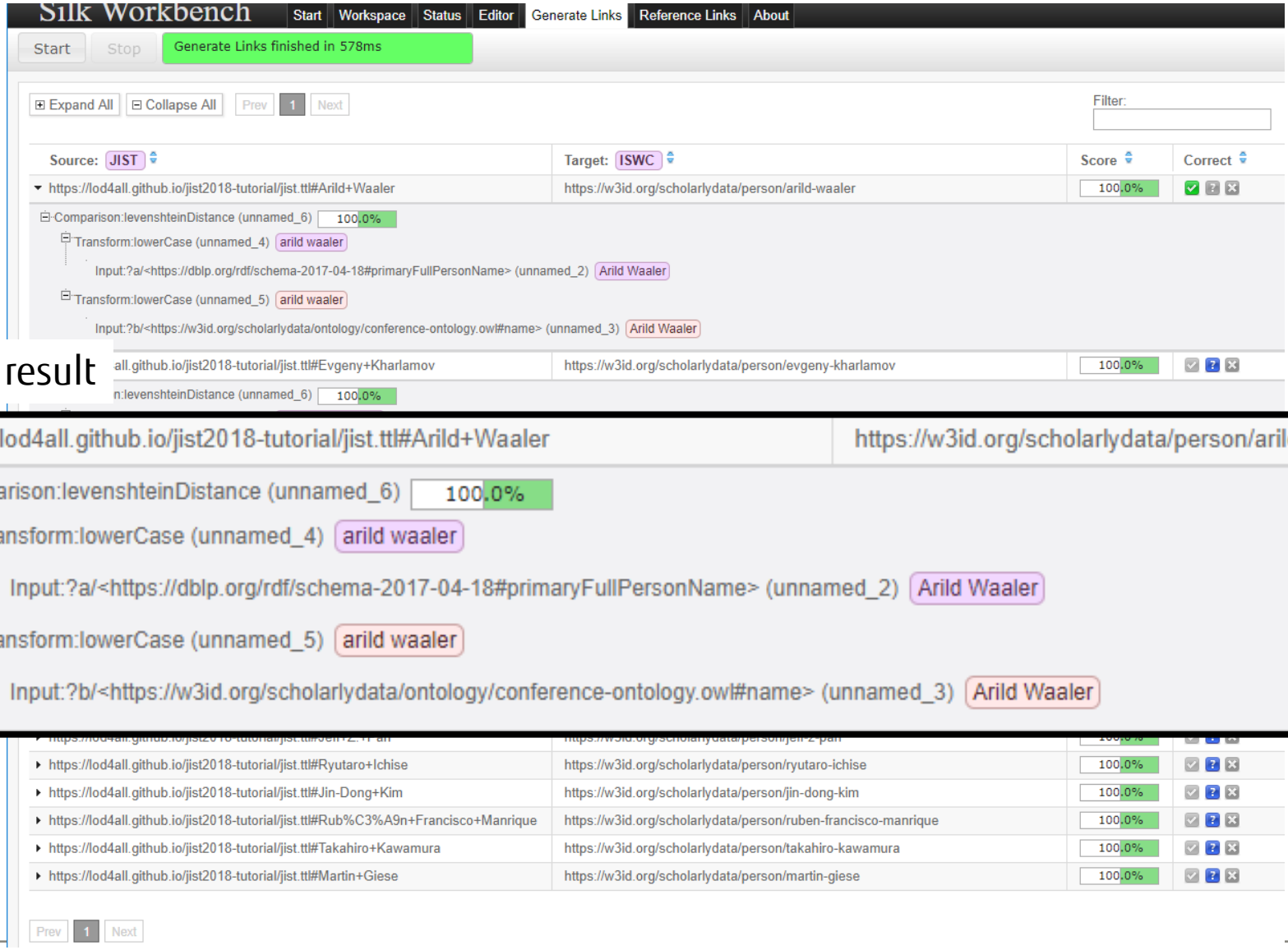
3. Calculate score with Levenstein distance



Generate Links: UI to run the linking task and check the results

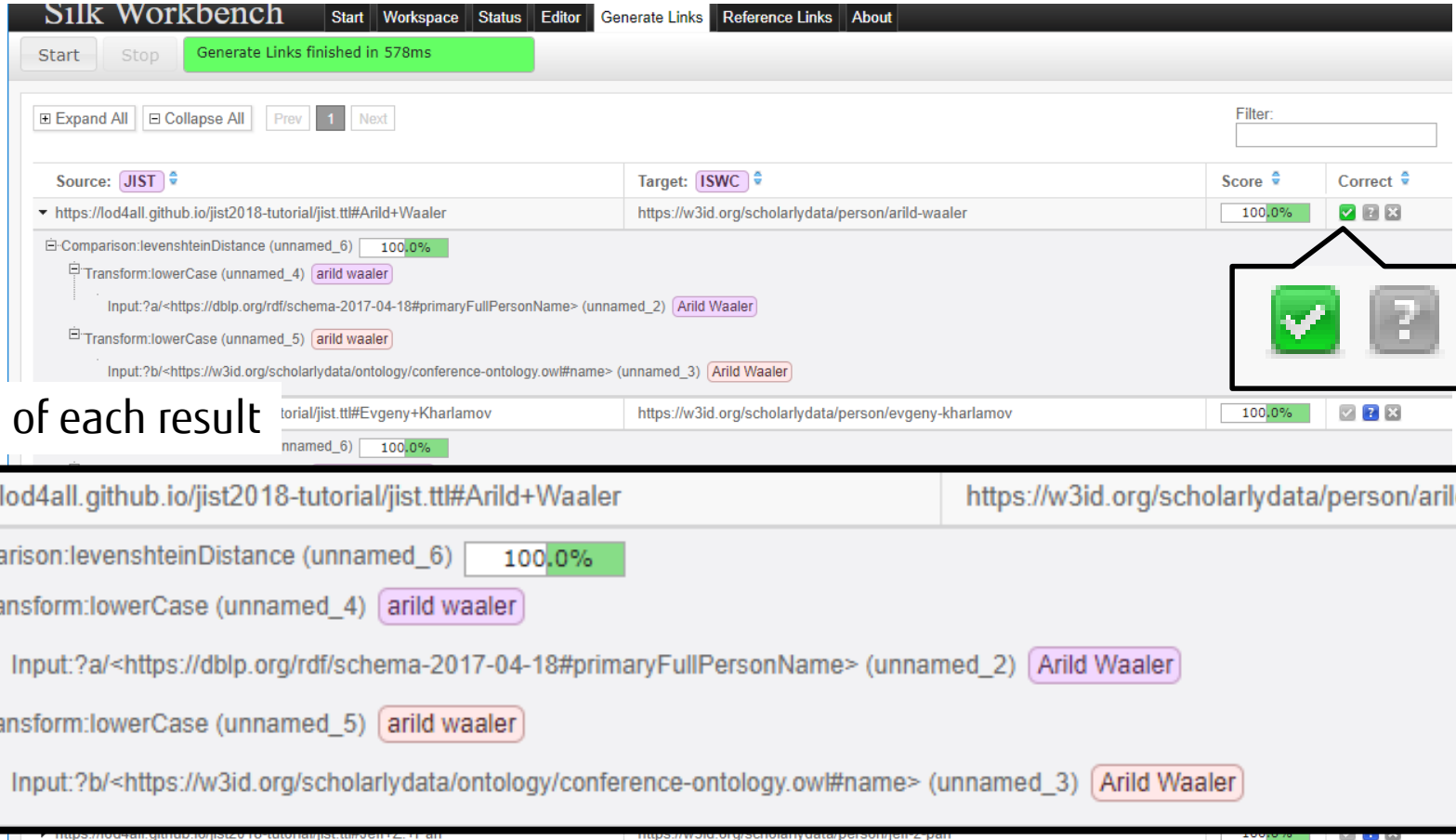


Generate Links: UI to run the linking task and check the results



Each result

Generate Links: UI to run the linking task and check the results

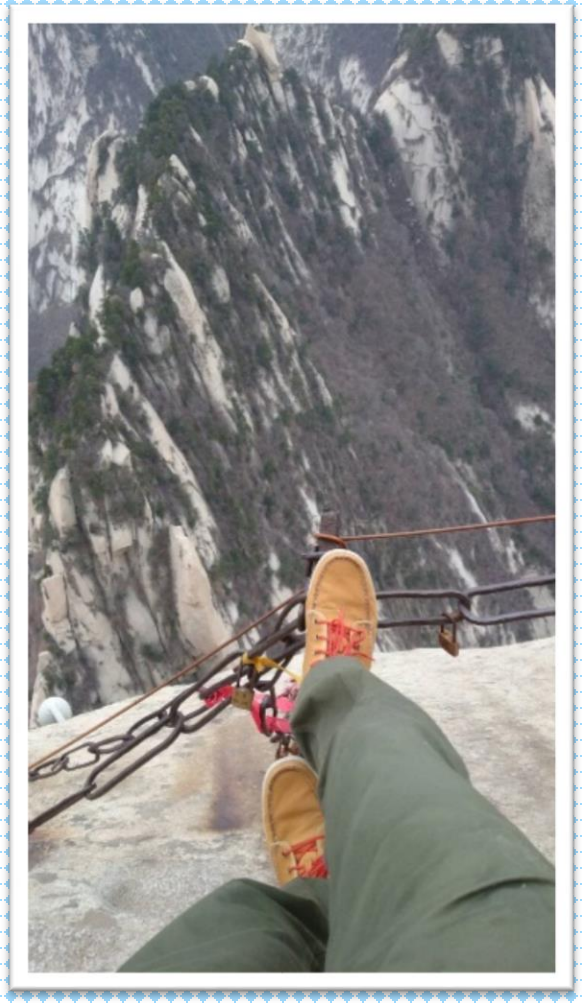


Detail of each result

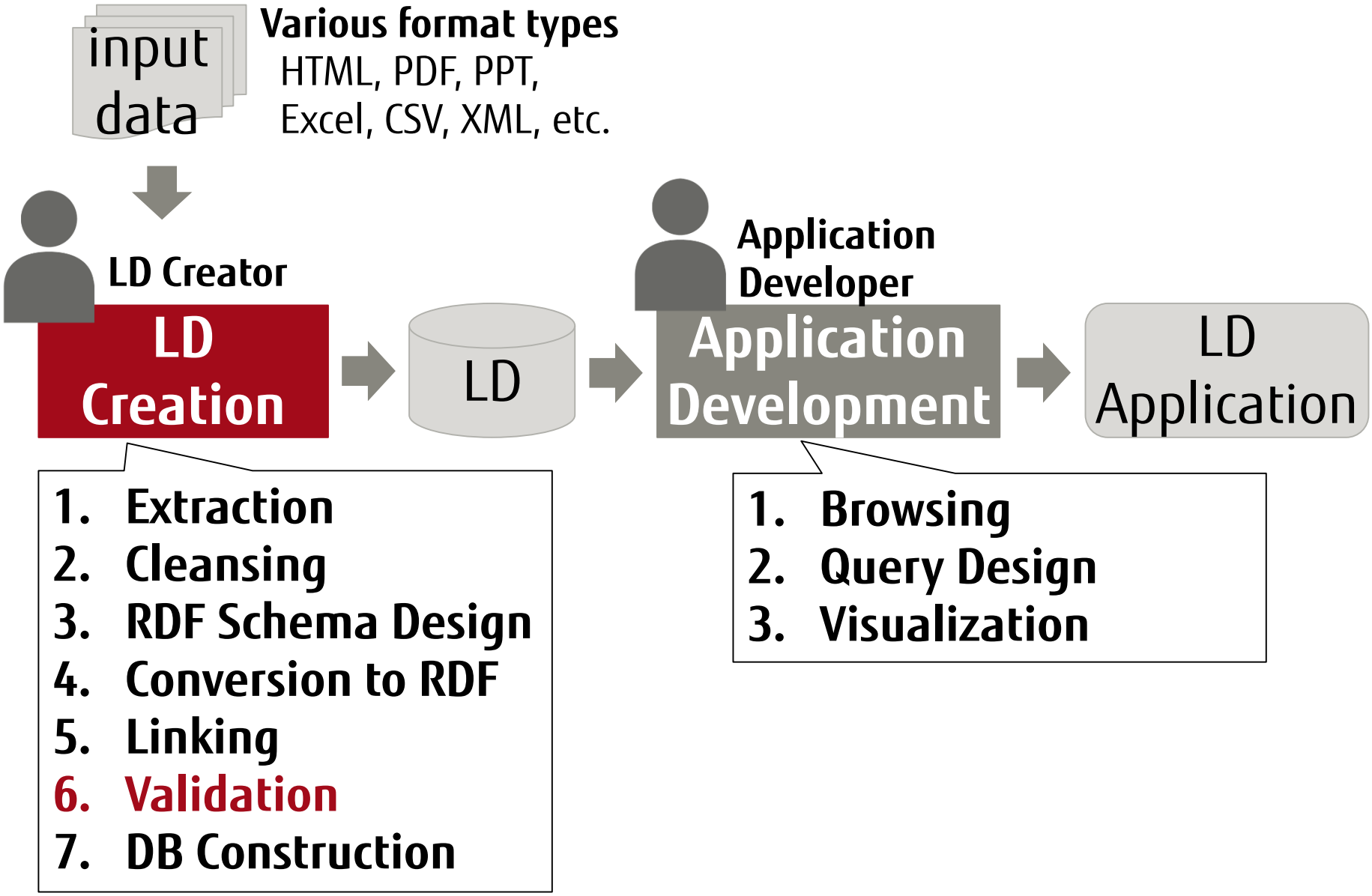
My Favorite Things

FUJITSU

Mount Hua / 華山 (China)



Outline



Validation

FUJITSU

To validate the syntax and the schema for the created RDF files

- Syntax validation
  - rapper★, Serd, arq
- Schema validation
  - Simple validation: SPARQL
  - Languages for validation: SHACL★, ShEx

rapper ★

Validation (1/6)

(Official) http://librdf.org/raptor/rapper.html

FUJITSU

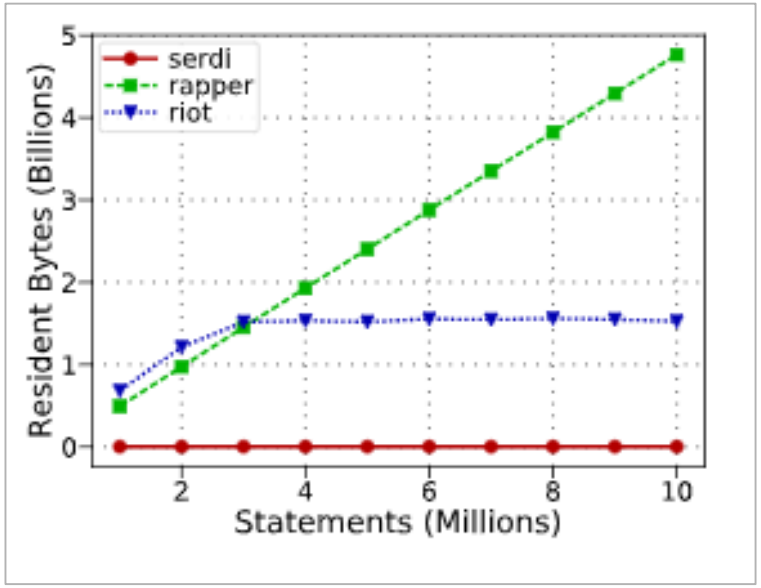
- A RDF parsing and serializing utility
  - Raptor RDF Syntax Library
- Usage
  - RDF serializing
    - rapper –output <format> –input <format> <inputfile>
  - GraphViz DOT Serializer
    - rapper --output dot --input <format> <file> | dot -Tpng -o output.png
  - **Validating RDF content** ★
    - rapper -c –input <format> <file>
- Problem
  - Official site version difficult to handle huge files
    *Latest version: 2.0.15
    *Github version improved this problem by reading input in chunks
    https://github.com/dajobe/raptor

Serd

https://drobilla.net/software/serd

FUJITSU

- A RDF parsing and serializing utility
  - Free Software released under the extremely liberal ISC license
- Features
  - It can handle huge files
    - Adopting stream processing
    - Comparing with rapper
- Usage
  - RDF serializing
    - serdi -i <format> -o <format> input_file
    - serdi -i <format> -o <format> -s input_string
    - format:turtle or Ntriples
    - STD output
- Others
  - Making URI of Blank node



- Command for executing SPARQL queries for RDF files
  - One of Jena package's command-line tools
  - arq is used for validating IRI (rapper CANNOT validate IRI)

- Usage
  - arq --data=<file> --query=<queryfile>
  - arq --data=<file> '<query>'

- Comparison of arq and rapper for IRI validation

arq

```
$ arq -data arq_vs_rapprt.ttl --query sample.rq
10:38:40 WARN  riot                 :: [line: 6, col:
1 ] Bad IRI: <http://bad://example.com#bob> Code:
12/PORT_SHOULD_NOT_BE_EMPTY in PORT: The colon
introducing an empty port component should be omitted
entirely, or a port number should be specified.
```

found Bad IRI
(has double slashes)

rapper

```
$ rapper -c --input turtle arq_vs_rapprt.ttl
rapper: Parsing URI file:///root/arq_vs_rapprt.ttl with
parser turtle
rapper: Parsing returned 3 triples
```

NOT found Bad IRI

SPARQL Query

FUJITSU

- It is a means w/o validation tools
- Invalid RDF Data

```
# cat jist_ap-json.include.invalidation.data.ttl | head -n 10
<https://lod4all.github.io/9> dc:identifier "9" , "99999";
```

✗ multiple IDs

- SPARQL Query

```
# cat sparql/rule.rq
CONSTRUCT { ?paper v:isValid false . }
WHERE{ FILTER(?number_of_id != 1)
  { SELECT ?paper (COUNT(DISTINCT ?id) AS ?number_of_id)
    WHERE {
      ?paper a  dblp:Publication ;
             dc:identifier ?id .
    } GROUP BY ?paper
  }
}
```

check number of IDs

- Result

```
# arq -data jist_ap-json.invalidation.data.ttl -query sparql/rule.rq
<https://lod4all.github.io/9> v:isValid  false .
```

- Describes structures (constraints) of an RDF graph as Shapes
  - ex)：a resource has (or doesn't have) specific predicates .
- Usage

Shape file

```
sh:rule1
  a sh:NodeShape ;
  sh:targetClass dblp:Publication ;
  sh:property [
    sh:path dc:identifier ;
    sh:minCount 1 ;
    sh:maxCount 1 ;
    sh:nodekind xsd:string;
  ] .
```

constraints for number of IDs

Result

```
# pyshacl -s shacl/rule.ttl -d jist_ap-json-20181101.ttl
Validation Report
Conforms: False
Results (1):
Constraint Violation in MaxCountConstraintComponent
    (http://www.w3.org/ns/shacl#MaxCountConstraintComponent):
    Severity: sh:Violation
    Source Shape: [ sh:maxCount Literal("1", datatype=xsd:integer) ; ... ]
    Focus Node: <https://lod4all.github.io/jist2018-tutorial/jist.ttl#9>
    Result Path: dc:identifier
```

- Language for constraints with equivalent function to SHACL
  - Compact notation is implemented unlike SHACL

- Comparison for Shape files

SHACL

```
sh:property [
    sh:path dc:identifier ;
    sh:minCount 1 ;
    sh:maxCount 1 ;
    sh:nodekind xsd:string;
] .
```

RDF syntax
(machine-readable)

ShEx

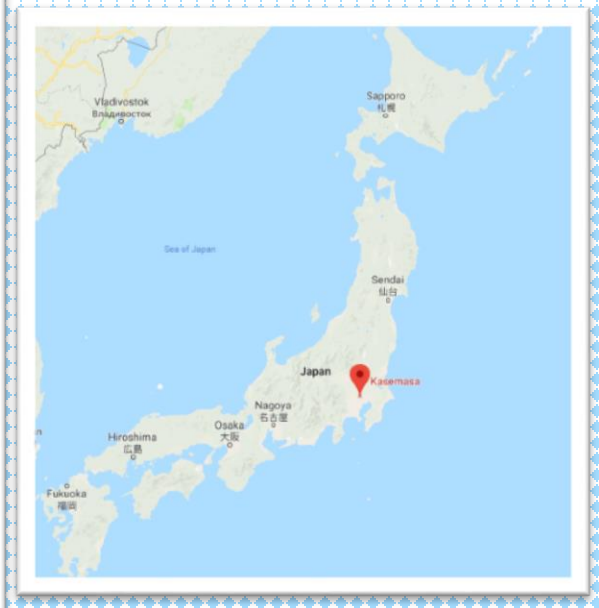
```
dc:identifier xsd:string{1}
```

ShexC syntax
(human-readable)

My Favorite Things

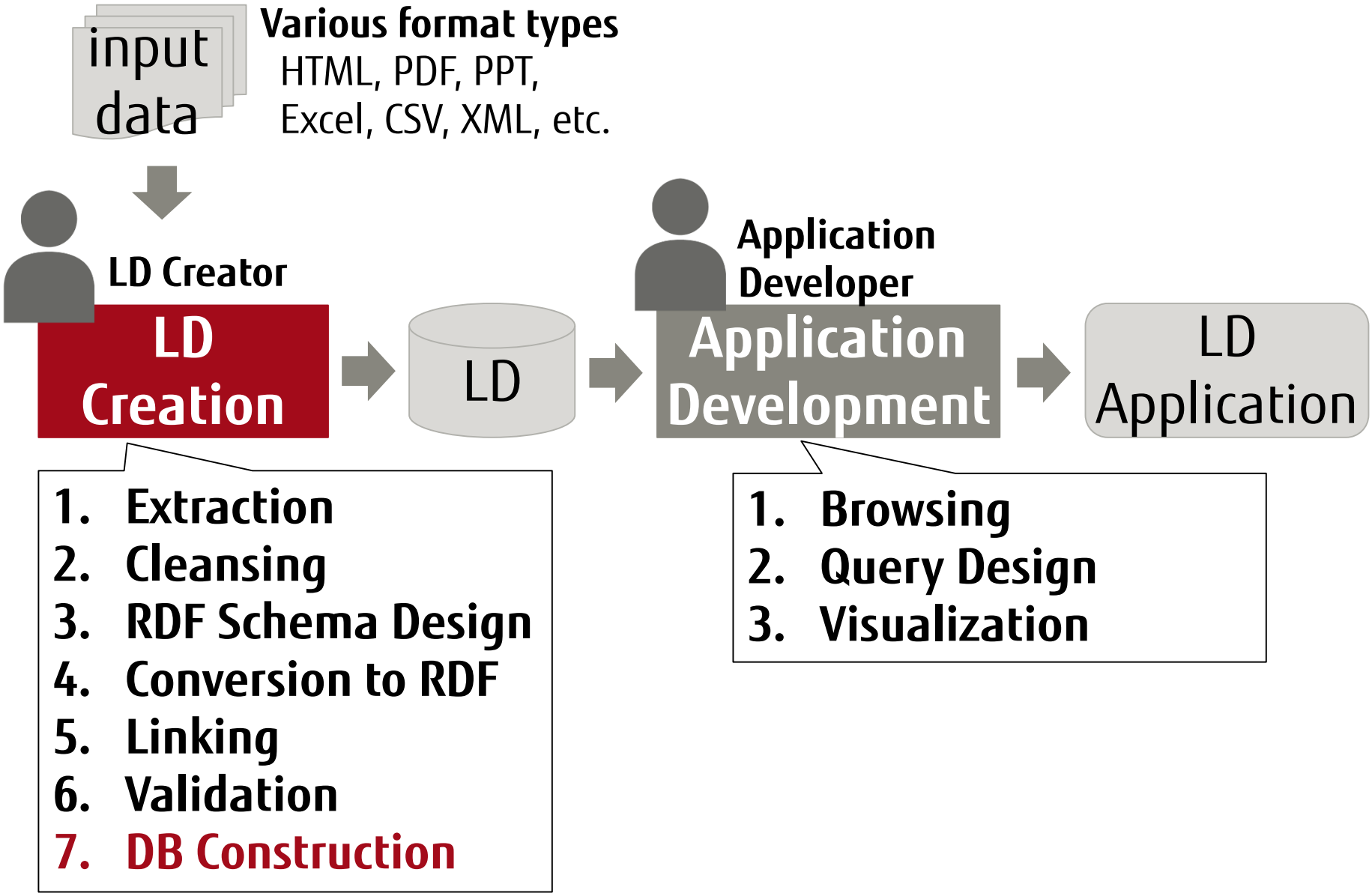
FUJITSU

Soft Roe Nabe/白子鍋



Kasemasa Bar / 加瀬政
Tokyo, Japan
https://tabelog.com/tokyo/A1323/A132301/13003798/

Outline

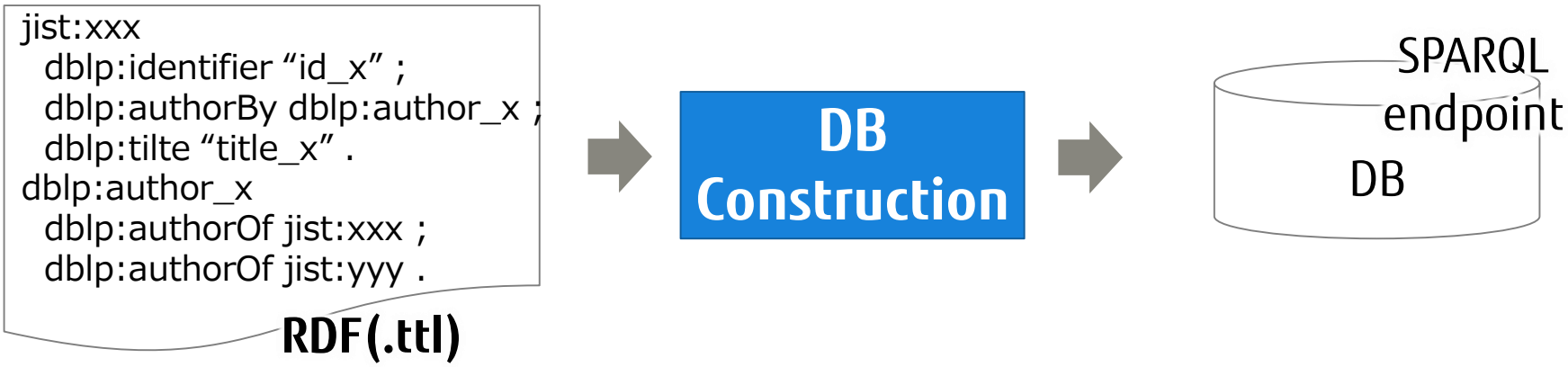


DB Construction

FUJITSU

Loading the dataset into the DB to enable users to access the SPARQL endpoint

Ex. JIST data

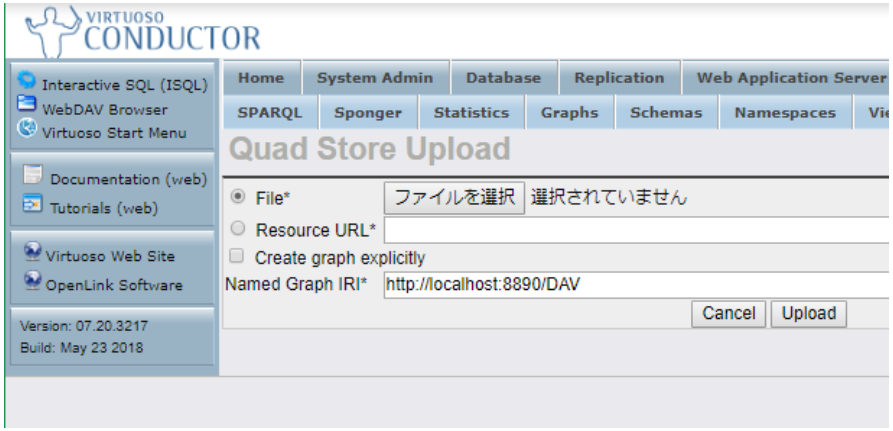


DB Construction (cont.)

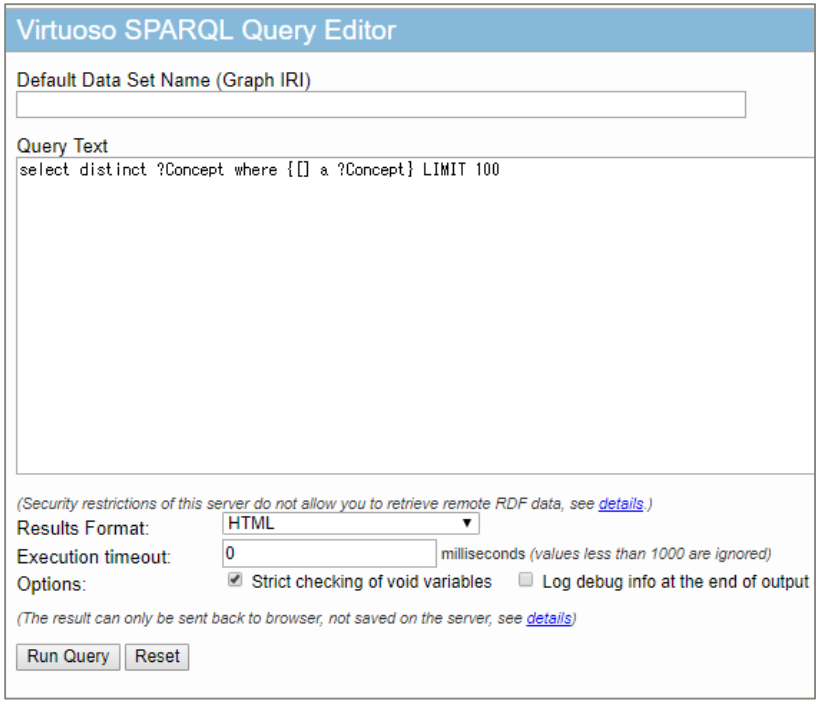
FUJITSU

- OpenLink Virtuoso★ https://virtuoso.openlinksw.com/
  - Latest version:07.20.3229
  - License: Single Server Edition: Free, Universal Server: Commercial license
  - Acquisition:
    - git clone -b stable/7 https://github.com/openlink/virtuoso-opensource
- StarDog https://www.stardog.com/
  - Latest version: Stardog 5.3.5 (3 Oct 2018)
  - Acquisition: Requiring user registration
- AllegroGraph https://franz.com/agraph/allegrograph/
  - Latest version : Release 6.4.4
  - Acquisition: https://franz.com/agraph/downloads/server?ui=new
- GraphDB http://graphdb.ontotext.com/documentation/standard/release-notes.html
  - Latest version : GraphDB 8.7.2 (22 October 2018)
  - Acquisition: https://www.ontotext.com/products/graphdb/
    - Requiring user registration

- High performance and large capacity RDF store
  - Single Server Edition:Free
  - Note: In case of clustering system, you need to contract a license



RDF loading UI

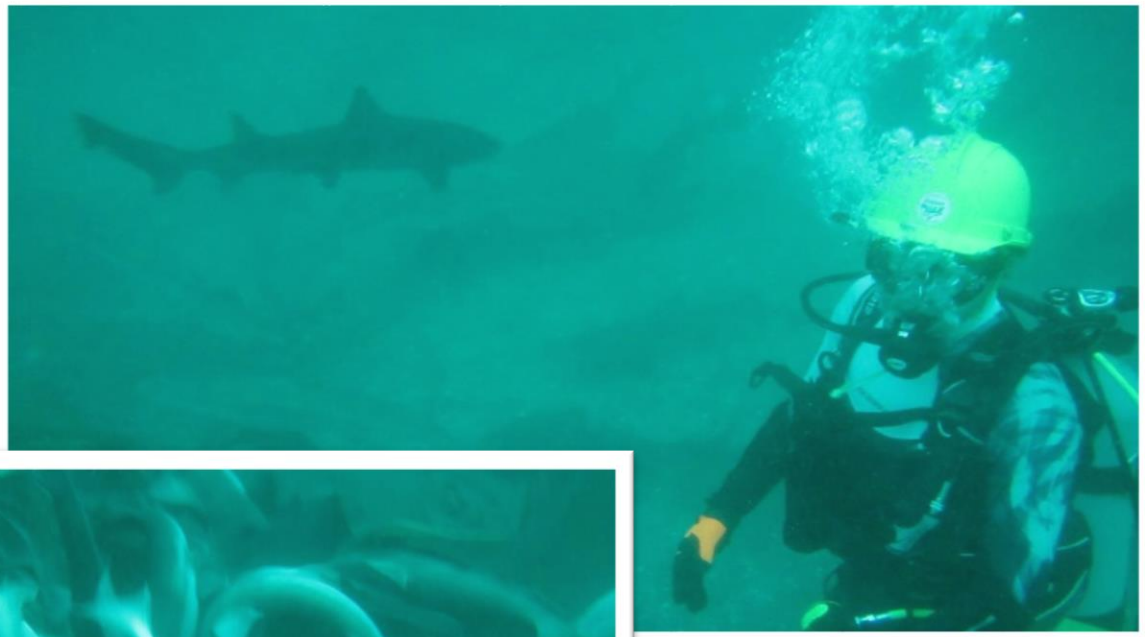


SPARQL endpoint UI

My Favorite Things

FUJITSU

Shark Diving



Chiba, Japan

Review of LD Creation

FUJITSU

Extraction ★Beautiful Soap

Cleasing ★Open Refine

RDF Schema Design ★Linked Open Vocabraries ★Google Dataset Search ★Data Sniffier

Conversion to RDF ★Open Refine +RDF Extension

Linking ★Silk

Validation ★rapper ★arq ★SHACL

DB Construction ★Virtuoso

RDF(.ttl)

```
jist:xxx
  dblp:identifier “id_x” ;
  dblp:authorBy dblp:author_x ;
  dblp:tilte “title_x” .
```

SPARQL endpoint

DB

Let's have a break

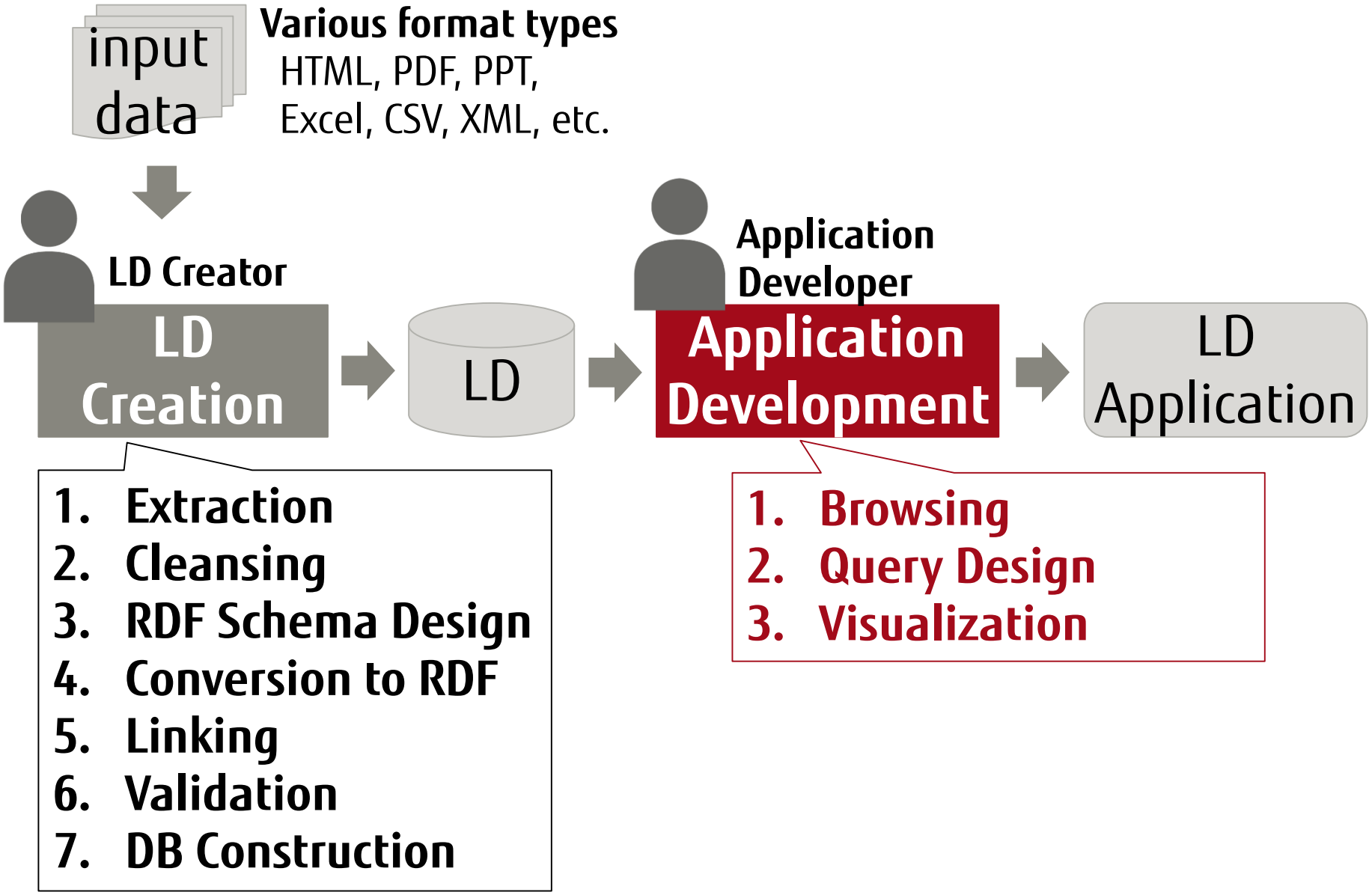
FUJITSU

Morning	Introduction to LD tools	9:00-10:00 (60 mins)	Presentation
		10:00-10:10 (10 mins)	Break
		10:10-10:50 (40 mins)	Presentation
		10:50-11:10 (20 mins)	Use Case Studies
Afternoon	Hands-on	13:00-13:10 (10 mins)	Introduction
		13:10-14:55 (95 mins)	Trial 1: Let's visualize corporate data Trial 2: Let's develop a financial analysis application
		14:55-15:00 (5 mins)	Closing

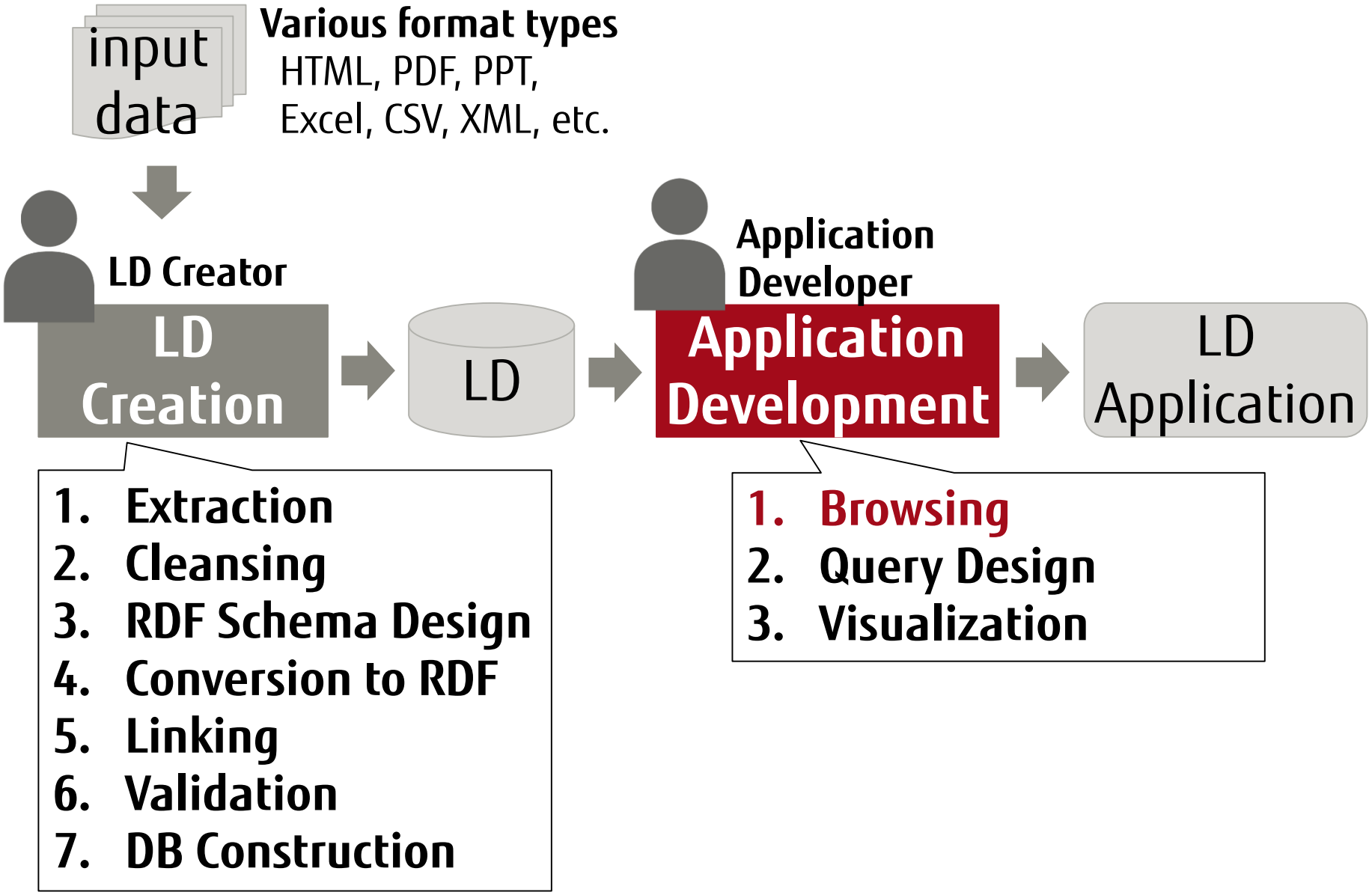
FUJITSU

2. Application Development

Outline



Outline



Browsing

FUJITSU

To browse dataset contents from SPARQL endpoints or RDF files directly

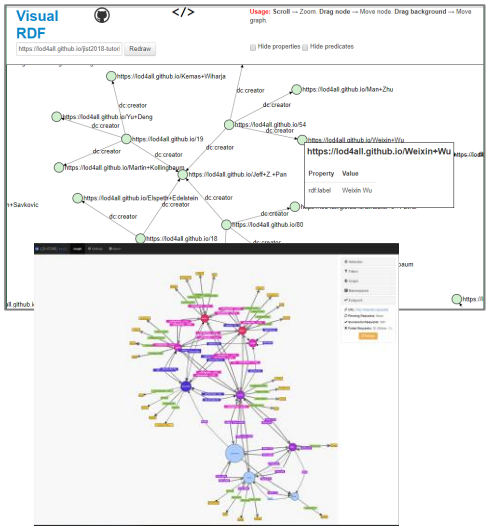
Ex. JIST data

```
jist:xxx
  dblp:identifier “id_x” ;
  dblp:authorBy dblp:author_x ;
  dblp:tilte “title_x” .
dblp:author_x
  dblp:authorOf jist:xxx ;
  dblp:authorOf jist:yyy
```

RDF(.ttl)

SPARQL Endpoint

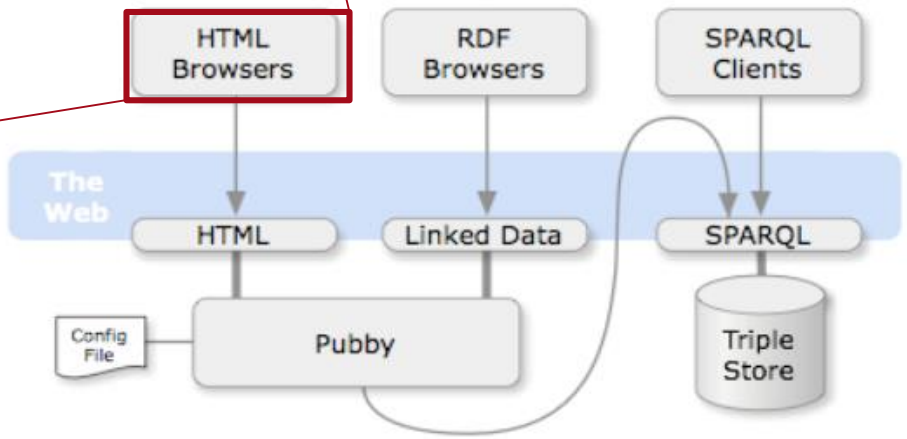
DB



- Table: Pubby★, Elda, DataSniffer, Loupe★
- Graph: ViaualRDF★, Graphviz★, Lodlive, LD-VOWL

- A Linked Data Frontend for SPARQL Endpoints
  - Mapping the original URIs to dereferenceable URIs
  - Asking information about the resource via URIs
    - handle requests to the mapped URIs by connecting to the SPARQL endpoint
  - Visualizing the response from the SPARQL endpoint as a table type view in one HTML page.

Table type view in HTML page



Elda LDA

https://www.epimorphics.com/technology/linked-data-api/

FUJITSU

- A linked data API for web developers to access linked data from a triple store
  - It provides a table type view for RDF data
  - The linked data can accessed both by SPARQL query, and directly from JavaScript via a simple, straightforward API.

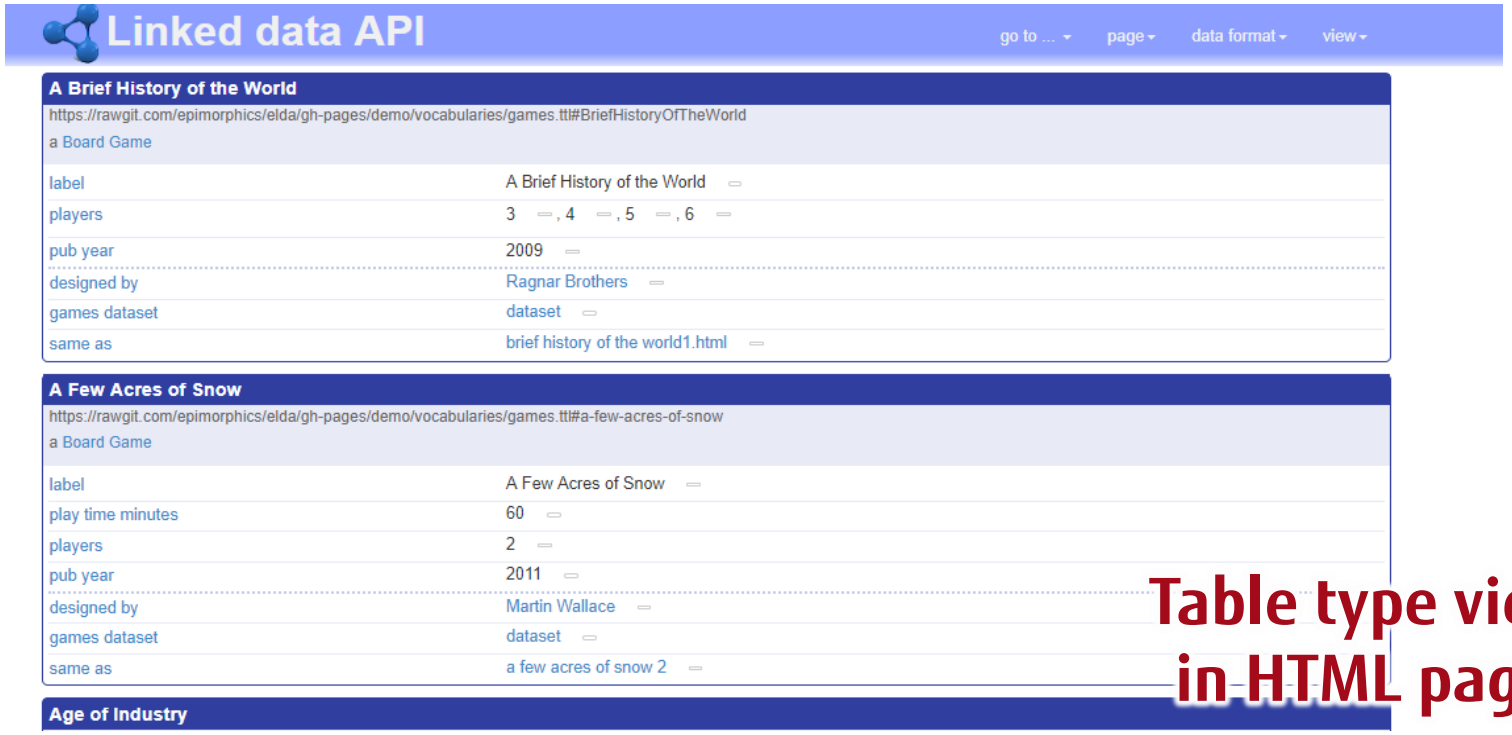


Table type view in HTML page

OpenLink Structured Data Sniffer

Browsing (3/8)

FUJITSU

- Enable to confirm RDF and metadata in HTML
  - Chrome extension

DBLP website

[–] Publication search results

found 36 matches

2017

Franz Baader, Daniel Borchmann, Adrian Nuradiansyah:
Identity Problem in Description Logic Ontologies and Its Application to View-Based Information
g. JIST 2017: 102-117

export record
BibTeX
RDF N-Triples
XML
dblp key:
conf/jist/BaaderBN17

Turtle

Statement Collection #1	
Entity	https://dblp.uni-trier.de/rec/conf/jist/BaaderBN17
Attributes	
owl:sameAs	https://dblp.org/rec/conf/jist/BaaderBN17
Statement Collection #2	
Entity	https://dblp.org/rec/conf/jist/BaaderBN17
Attributes	
rdf:type	https://dblp.uni-trier.de/rdf/schema-2017-04-18#Publication
owl:sameAs	https://doi.org/10.1007/978-3-319-70682-5_7
https://dblp.uni-trier.de/rdf/schema-2017-04-18#title	The Identity Problem in Description Logic Ontologies and Its Application to View-Based Information Hiding.
https://dblp.uni-trier.de/rdf/schema-2017-04-18#bibtexType	http://data.bibbase.org/ontology/#Inproceedings
https://dblp.uni-trier.de/rdf/schema-2017-04-18#publicationType	https://dblp.uni-trier.de/rdf/schema-2017-04-18#Inproceedings
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/b/Baader:Franz
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/b/Borchmann:Daniel
https://dblp.uni-trier.de/rdf/schema-2017-04-18#authoredBy	https://dblp.org/pers/n/Nuradiansyah:Adrian
https://dblp.uni-trier.de/rdf/schema-2017-04-18#primaryElectronicEdition	https://doi.org/10.1007/978-3-319-70682-5_7
https://dblp.uni-trier.de/rdf/schema-2017-04-18#listedOnTocPage	https://dblp.org/db/conf/jist/jist2017
https://dblp.uni-trier.de/rdf/schema-2017-04-18#publishedInBook	JIST
https://dblp.uni-trier.de/rdf/schema-2017-04-18#pageNumbers	102-117
https://dblp.uni-trier.de/rdf/schema-2017-04-18#yearOfPublication	2017
https://dblp.uni-trier.de/rdf/schema-2017-04-18#publishedAsPartOf	https://dblp.org/rec/conf/jist/2017
Statement Collection #3	
Entity	https://dblp.org/rec/conf/jist/BaaderBN17.nt

- A tool that obtains statistics for each dataset
  - Searchable element
    - class / property(predicate) / triple pattern / named graph

Ex. Class statistic

List of used classes found in the **DBpedia (English)** dataset (**941** classes)

Show 10 entries

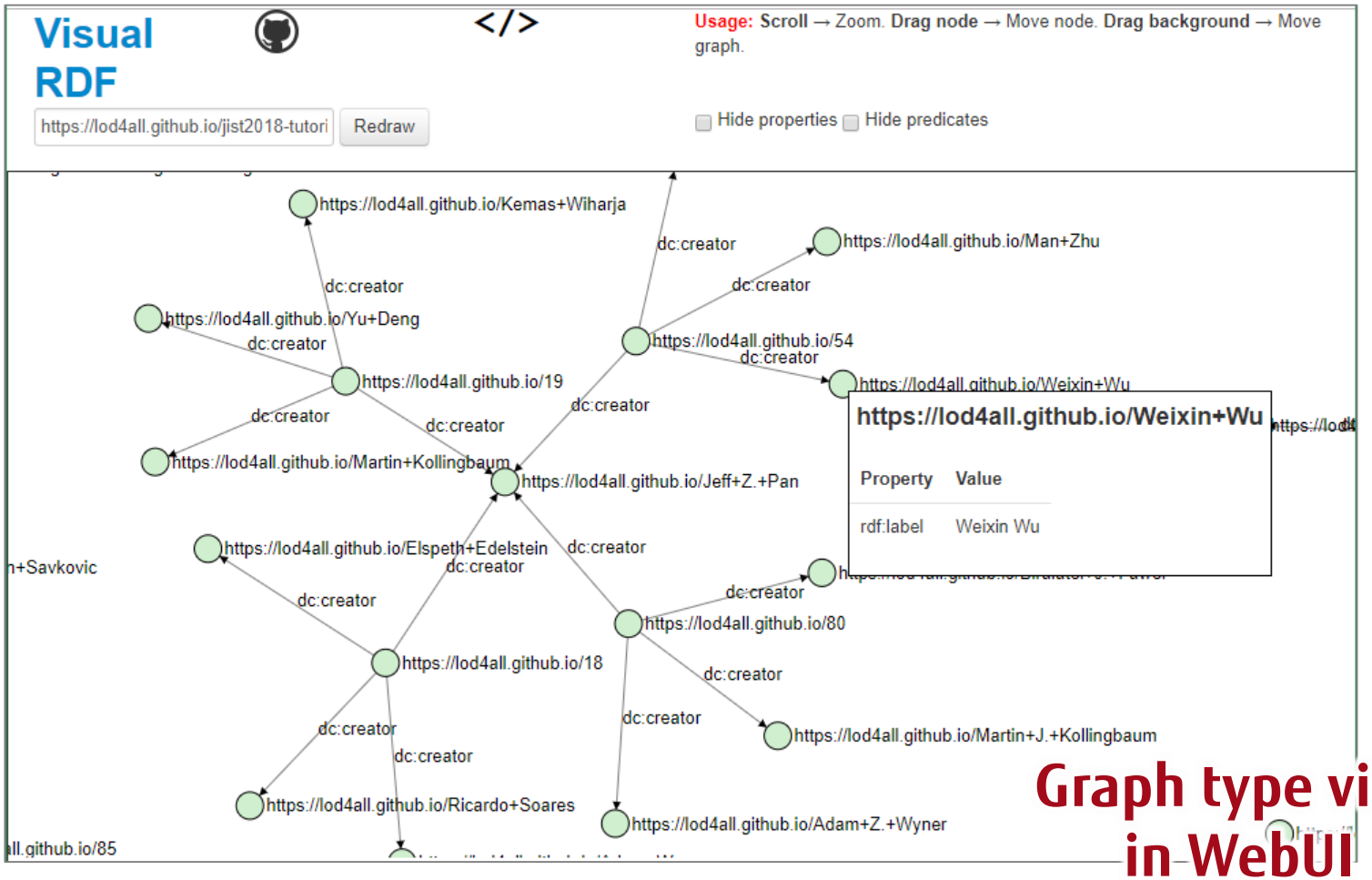
Class	Instance Count
http://dbpedia.org/ontology/Person	2135040
http://www.wikidata.org/entity/Q5	2135040
http://dbpedia.org/ontology/Organisation	220219
http://dbpedia.org/ontology/MusicalWork	162397

Count of Instances for each Class

Table type view for statistics

- Web Service that describe graph from a RDF file easily

Ex. JIST2018 accepted papers



Graph type view in WebUI

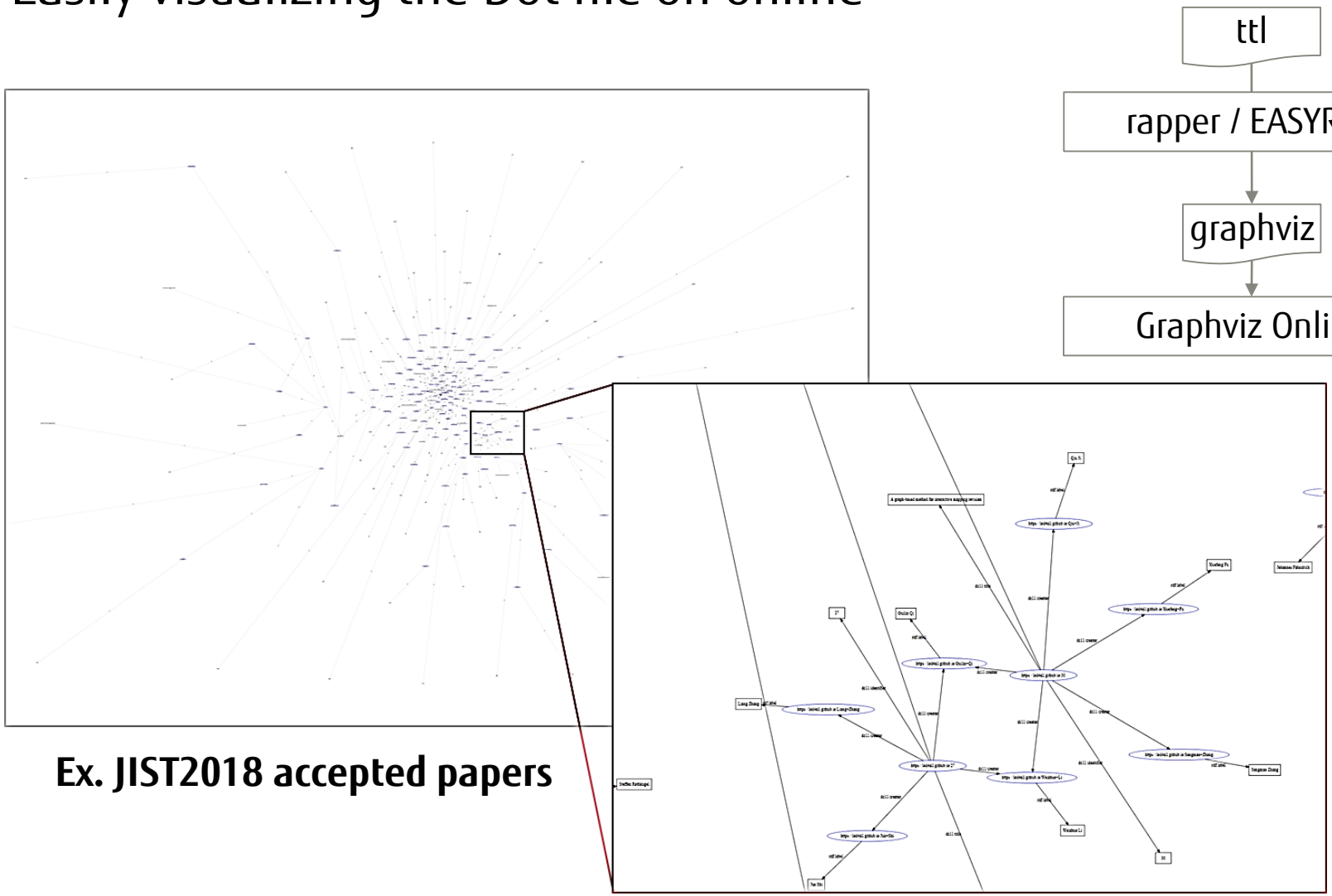
Graphviz Online ★

Browsing (6/8)

https://dreampuf.github.io/GraphvizOnline

FUJITSU

- Easily visualizing the Dot file on online



Ex. JIST2018 accepted papers

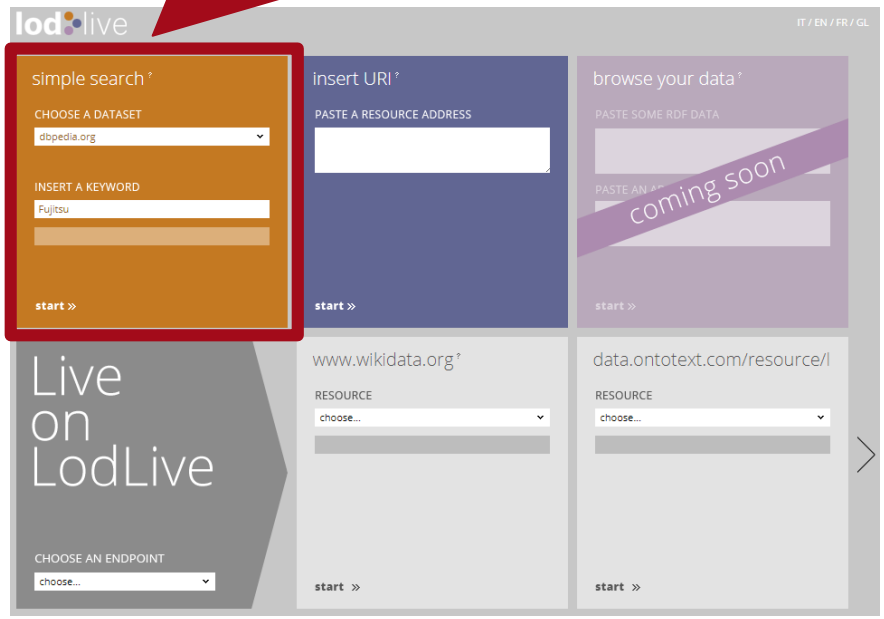
Lodlive

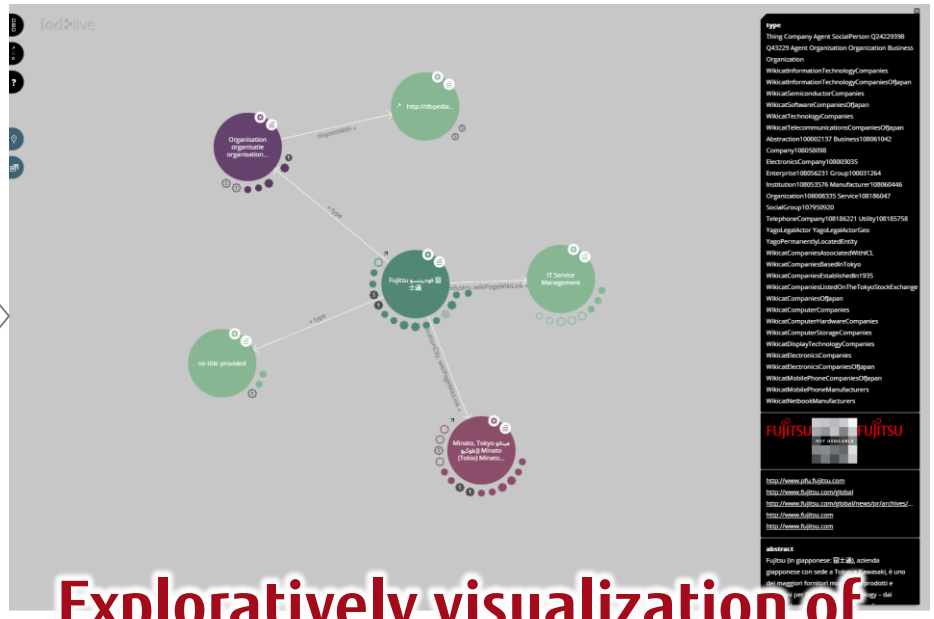
http://en.lodlive.it/

FUJITSU

- A web application for LOD visualization
  - It is developed with JavaScript and can be embedded in other applications

- Selecting a SPARQL endpoint
- Keyword search for interesting class





Exploratively visualization of the LOD from a selected class

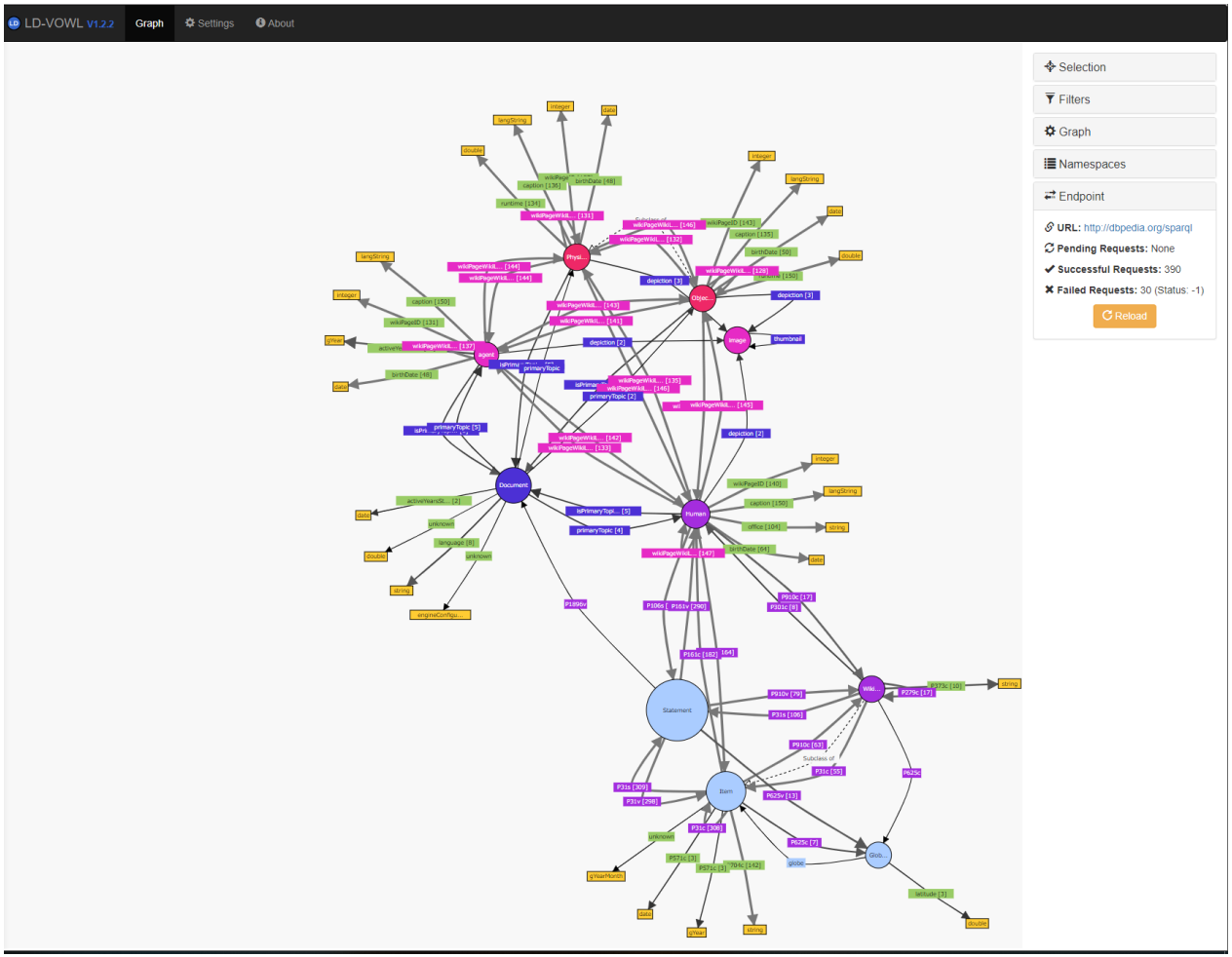
LD-VOWL

Browsing (8/8)

http://vowl.visualdataweb.org/ldvowl.html

FUJITSU

- Steamily extract the RDF data schemas from a specified SPARQL endpoint

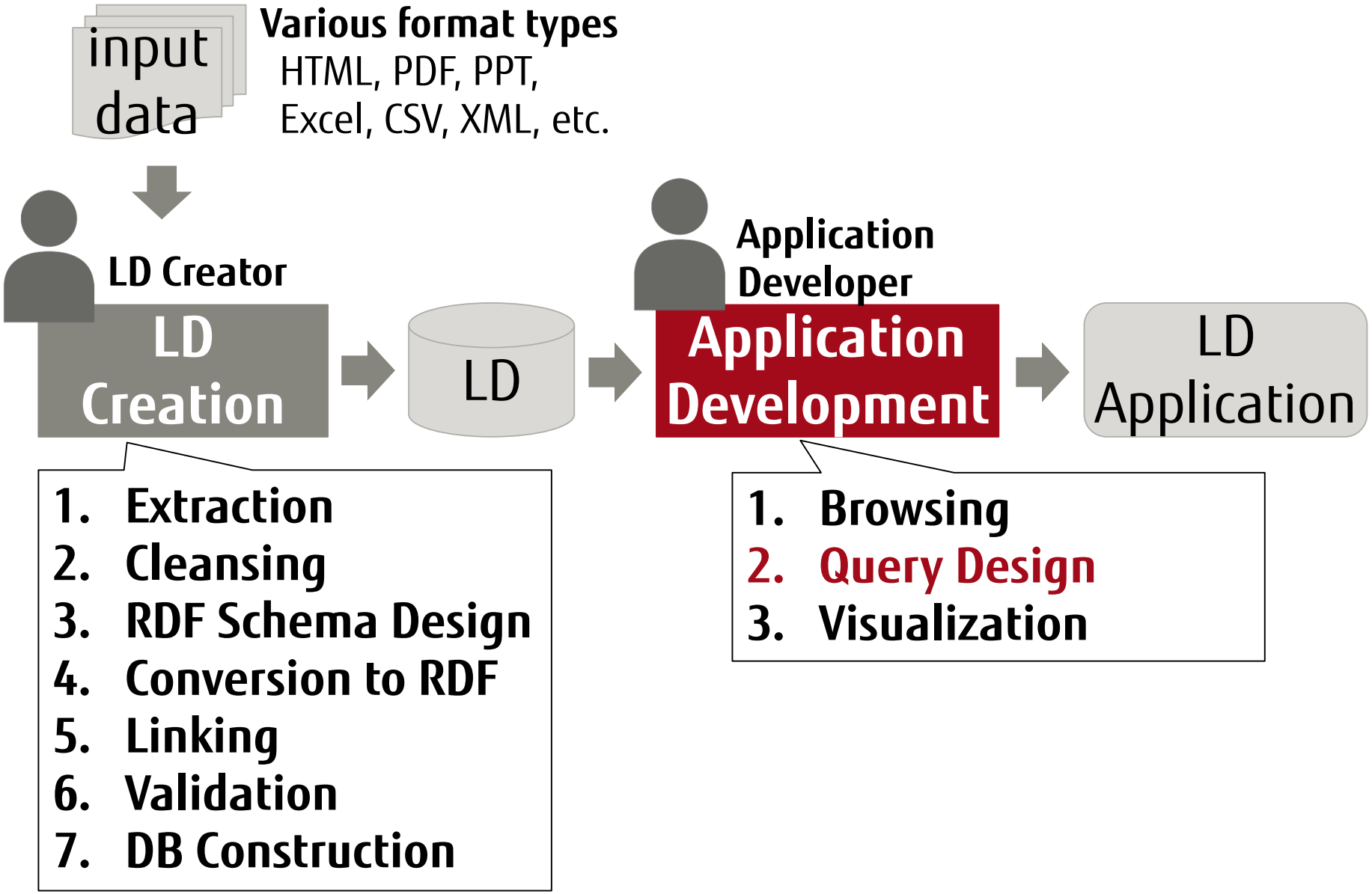


My Favorite Things

FUJITSU

Soft Shelld Turtle/すっぽん

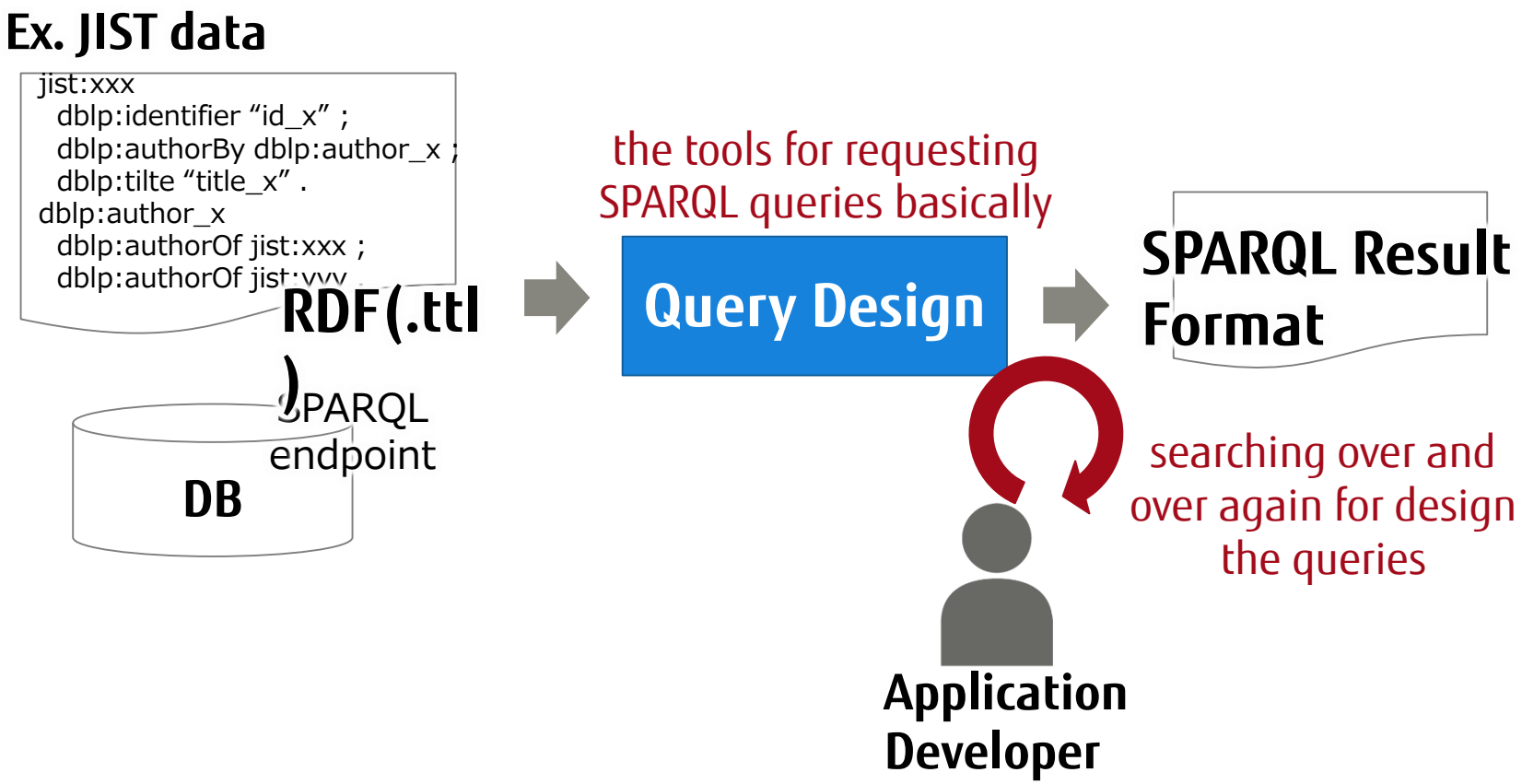
Outline



Query Design

FUJITSU

To design queries that search the necessary data for applications



Query Design (cont.)

FUJITSU

- for SPARQL endpoints
  - Command: rsparql★ SPANG
  - WebUI: sparql-kernel
  - library: SPARQL Wrapper(Python), SPARQL(R)
  - library independence: curl, Excel
- for RDF files
  - command: arq
  - library: rdflib

- Command-line tool to execute SPARQL for **SPARQL endpoint**
  - One of Jena packages
- Usage
  - rsparql –service=<sparql_endpoint_uri> –query=<queryfile>

```
PREFIX dblp: <https://dblp.org/rdf/schema-2017-04-18#>
SELECT ?title WHERE {
  ?paper dblp:title ?title;
    dblp:publishedInBook ?book.
  FILTER(regex(?book,"JIST"))
}
```

queryfile (query.rq)

```
$ rsparql --service http://[redacted] --query query.rq
```

Result

```
| title                                                                    |
==========================================================================
| "G-Diff: A Grouping Algorithm for RDF Change Detection on MapReduce."    |
| "A MapReduce-Based Approach for Prefix-Based Labeling of Large XML Data." |
| "Gathering Photos from Social Networks Using Semantic Technologies."     |
| "Publishing Danish Agricultural Government Data as Semantic Web Data."   |
```

- Command-line tool for executing queries to SPARQL endpoint
  - Specify endpoint and prefix in advance
- Usage

~/.spang/endpoints

```
dbpedia http://dbpedia.org/sparql
jisttutorial http://xxxxxxx/sparql
```

w/ queryfile

```
$ spang jisttutorial query.rq
"G-Diff: A Grouping Algorithm for RDF Change Detection on MapReduce."
"A MapReduce-Based Approach for Prefix-Based Labeling of Large XML Data."
…
```

~/.spang/prefix

```
PREFIX dblp: <https://dblp.org/rdf/schema-2017-04-18#>
```

```
$ spang jisttutorial -P dblp:title | cut -f2
"Integrated semantic model for complex disease network (Short Paper)"
"Implementing LOD Surfer as a Search System for the Annotation of
Multiple Protein Sequence Alignment (Short Paper)"
…
```

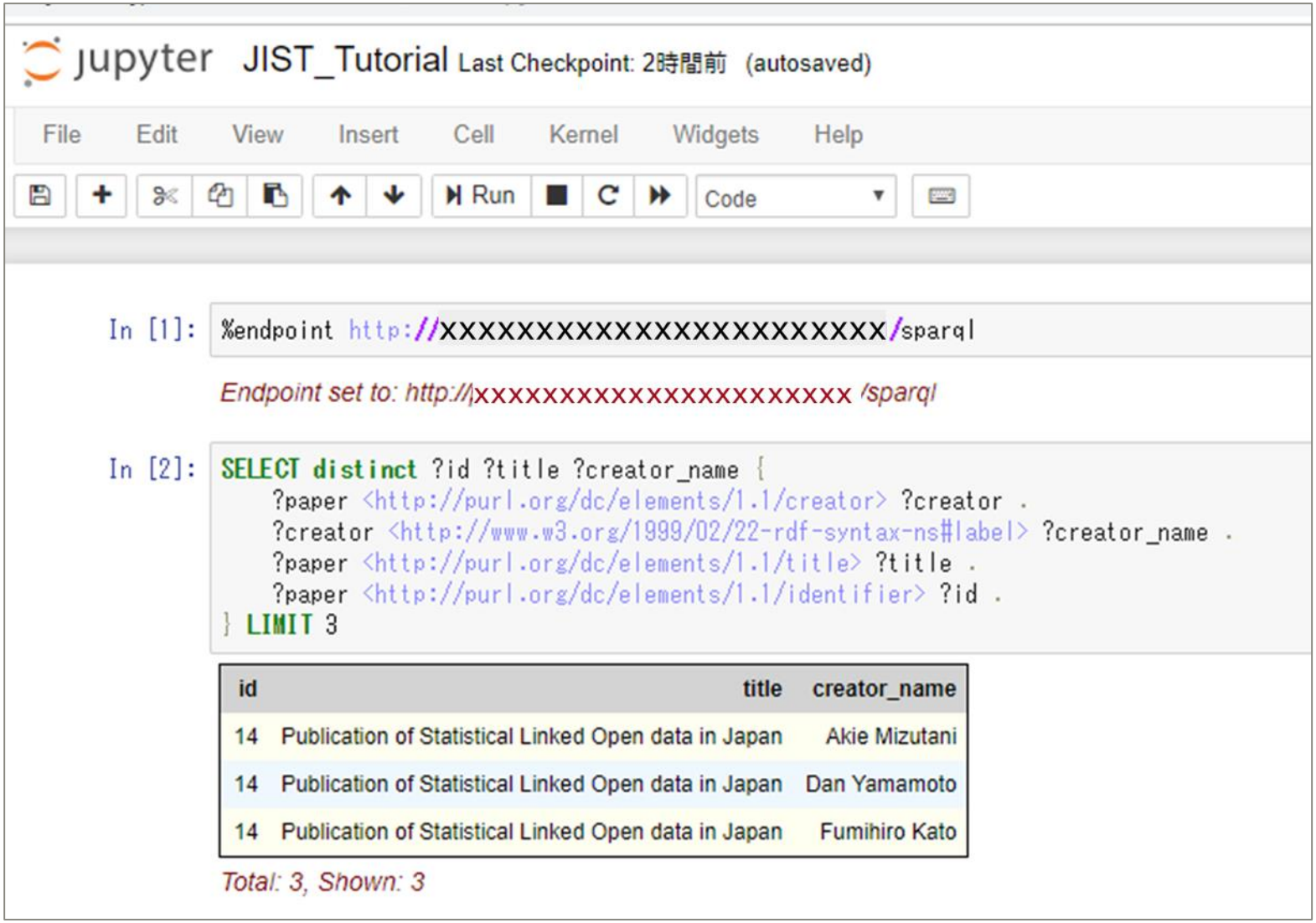
w/o queryfile

SPARQL kernel

https://github.com/paulovn/sparql-kernel

FUJITSU

- A Jupyter kernel for SPARQL

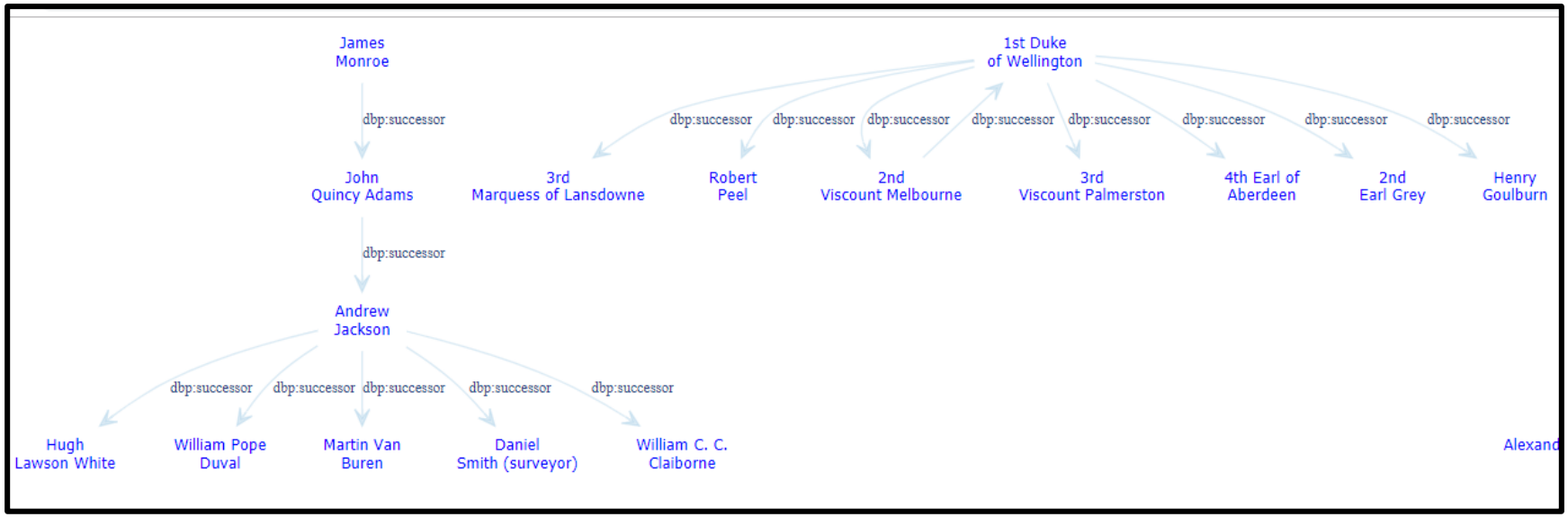


SPARQL kernel (cont.)

https://github.com/paulovn/sparql-kernel

FUJITSU

- It can output SVG in conjunction with GraphViz.
  - GraphViz must be installed on the same machine.



SPARQL Wrapper

https://rdflib.github.io/sparqlwrapper/

FUJITSU

- SPARQL endpoint interface for **Python library**
  - converts the result into various formats

Installation

```
$ pip install SPARQLWrapper
```

Code

```
from SPARQLWrapper import SPARQLWrapper, JSON

sparql = SPARQLWrapper("http://******/sparql")

sparql.setQuery("""
    select * where {
        ?paper dc:creator ?creator .
    } limit 3
""")
sparql.setReturnFormat(JSON)
print(sparql.query().convert())
```

SPARQL package for R

http://cran.r-project.org

FUJITSU

- R also has a SPARQL library

```
> library(SPARQL)
> endpoint <- http://******/sparql'
> q <- “PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT * WHERE {
        ?paper dc:creator ?creator .
> res <- SPARQL(endpoint, q)$results
```

- Command-line tool for transferring data

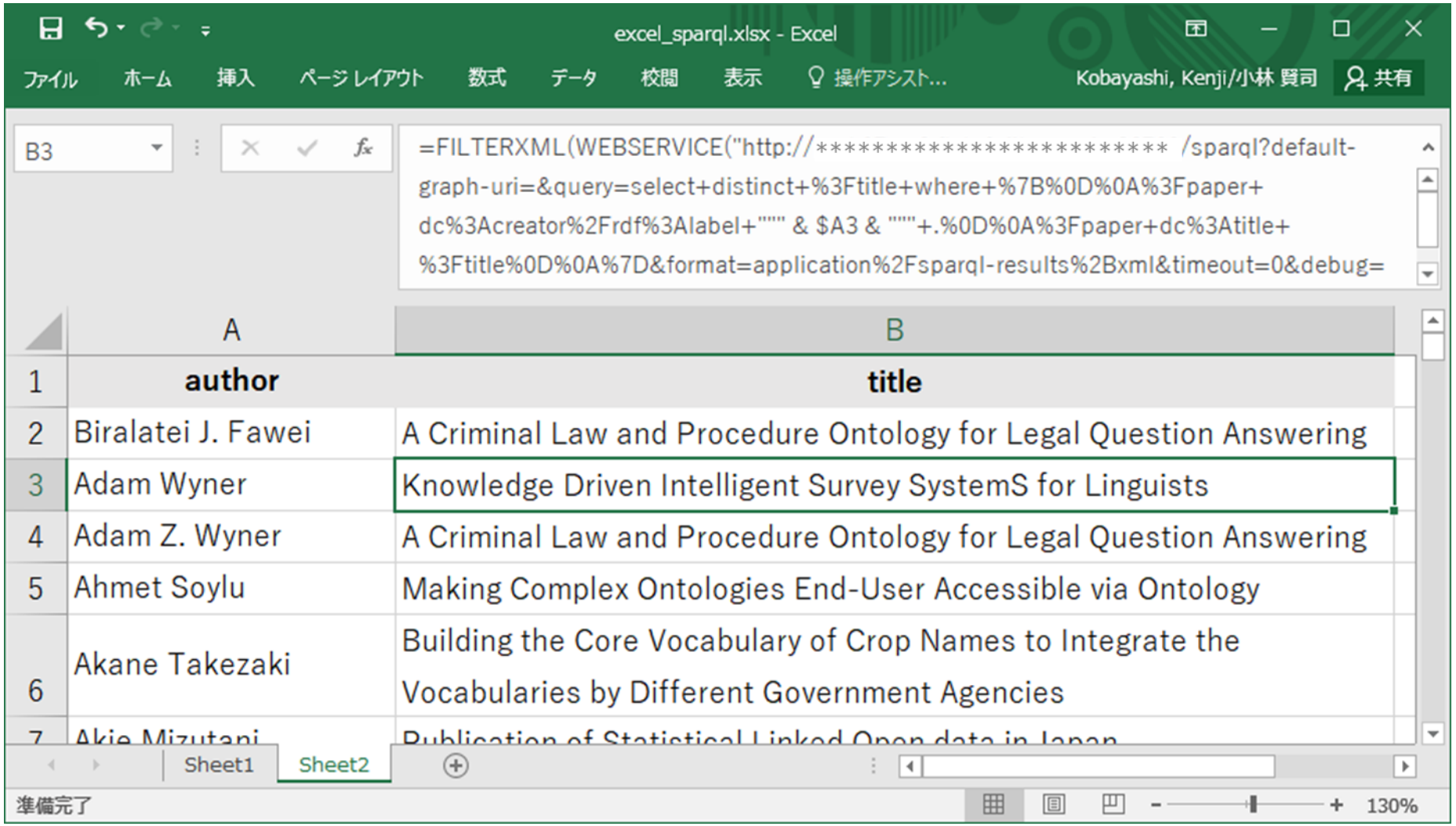
```
$ curl -G -H 'Accept: application/sparql-results+json' --data-urlencode 'query@-'
http://******/sparql < query_curl.rq


{ "head": { "link": [], "vars": ["creator"] },
  "results": { "distinct": false, "ordered": true, "bindings": [
    { "creator": { "type": "uri", "value": "https://lod4all.github.io/Akie+Mizutani" }},
    { "creator": { "type": "uri", "value": "https://lod4all.github.io/Dan+Yamamoto" }},
    { "creator": { "type": "uri", "value": "https://lod4all.github.io/Fumihiro+Kato" }},
    { "creator": { "type": "uri", "value": "https://lod4all.github.io/Hideaki+Takeda" }},
    { "creator": { "type": "uri", "value":
"https://lod4all.github.io/Hiromu+Harada" }} ] } }
```

Excel

FUJITSU

Use the WEBSERVICE function



	A	B
1	**author**	**title**
2	Biralatei J. Fawei	A Criminal Law and Procedure Ontology for Legal Question Answering
3	Adam Wyner	Knowledge Driven Intelligent Survey SystemS for Linguists
4	Adam Z. Wyner	A Criminal Law and Procedure Ontology for Legal Question Answering
5	Ahmet Soylu	Making Complex Ontologies End-User Accessible via Ontology
6	Akane Takezaki	Building the Core Vocabulary of Crop Names to Integrate the Vocabularies by Different Government Agencies
7	Akie Mizutani	Publication of Statistical Linked Open data in Japan

- Command line tool to execute SPARQL for **RDF dump**
  - One of Jena packages
- Usage
  - arq –data=<file> –query=<queryfile>
  - arq –data=<file> '<query>'

query.rq

```
SELECT distinct ?id ?title ?creator_name {
    ?paper <http://purl.org/dc/elements/1.1/creator> ?creator .
    ?creator <http://www.w3.org/1999/02/22-rdf-syntax-ns#label> ?creator_name .
    ?paper <http://purl.org/dc/elements/1.1/title> ?title .
    ?paper <http://purl.org/dc/elements/1.1/identifier> ?id .
} LIMIT 3
```

```
$ arq --data jist_ap-json.ttl --query query.rq
-----------------------------------------------------------------------------------------------------------------------
| id   | title                                                                                          | creator_name      |
=======================================================================================================================
| "48" | "Construction and Reuse of Linked Agriculture Data: An Experience of Taiwan Government Open Data" | "Dongpo Deng"     |
| "26" | "IMI: A Common Vocabulary Framework for Open Government Data"                                  | "Shuichi Tashiro" |
| "91" | "On Enhancing Visual Query Building Over KGs Using Query Logs (Short Paper)"                   | "Ahmet Soylu"     |
-----------------------------------------------------------------------------------------------------------------------
```

- Python library for RDF dump
  - specifies resource URIs or RDF files
  - used in SPARQL Wrapper library

Installation

```
$ pip install rdflib
```

Code

```
import rdflib
graph=rdflib.Graph()
graph.load("https://******/jist_ap-json.ttl",format="turtle")
for s,p,o in graph:
    print(s,p,o)
```

RDF file

python collection

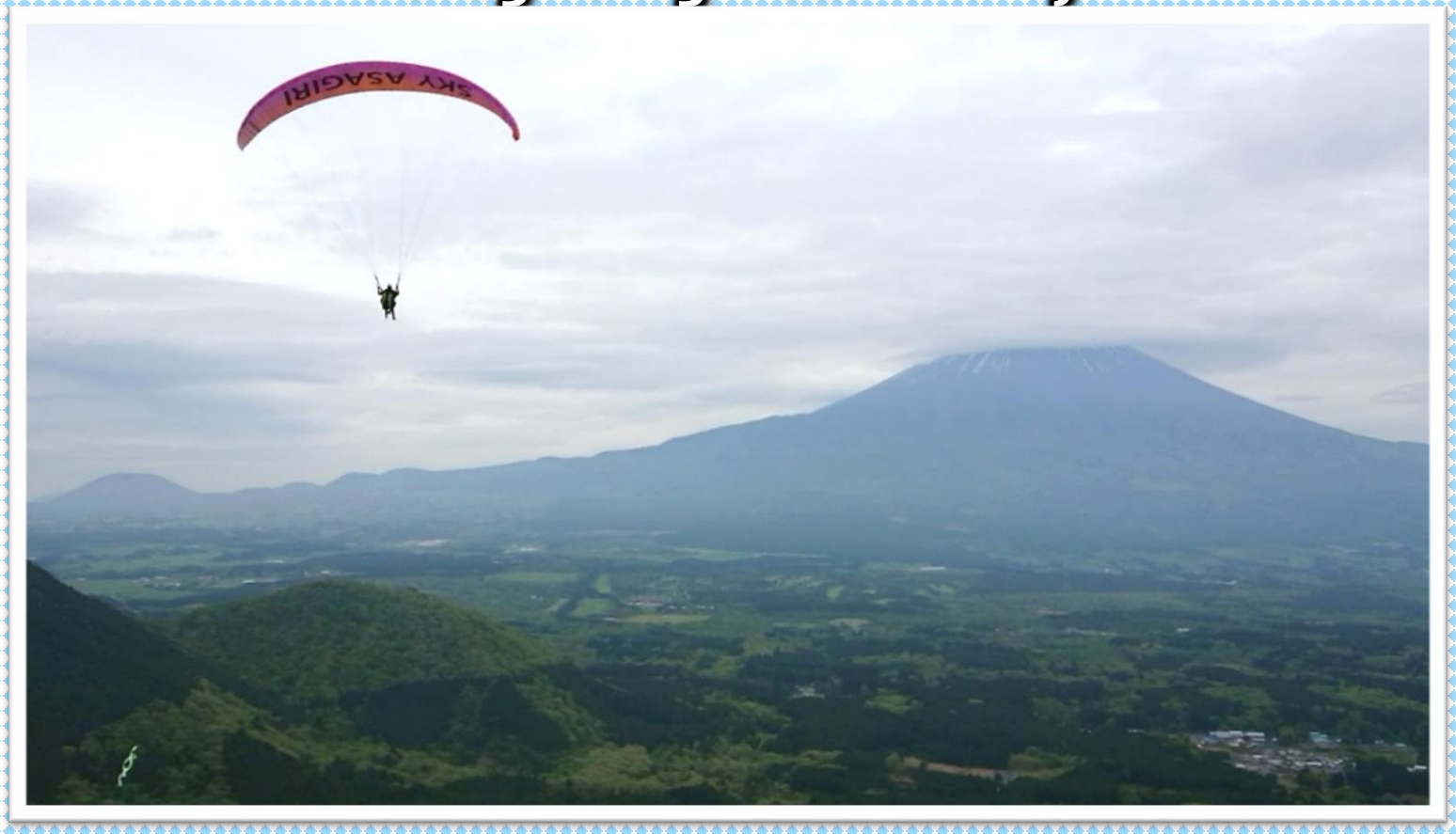
Result

```
https://lod4all.github.io/43 http://purl.org/dc/elements/1.1/creator https://lod4all.github.io/Pokpong+Songmuang
https://lod4all.github.io/30 http://purl.org/dc/elements/1.1/creator https://lod4all.github.io/Songmao+Zhang
https://lod4all.github.io/Khai+Nguyen http://www.w3.org/1999/02/22-rdf-syntax-ns#label Khai Nguyen
```

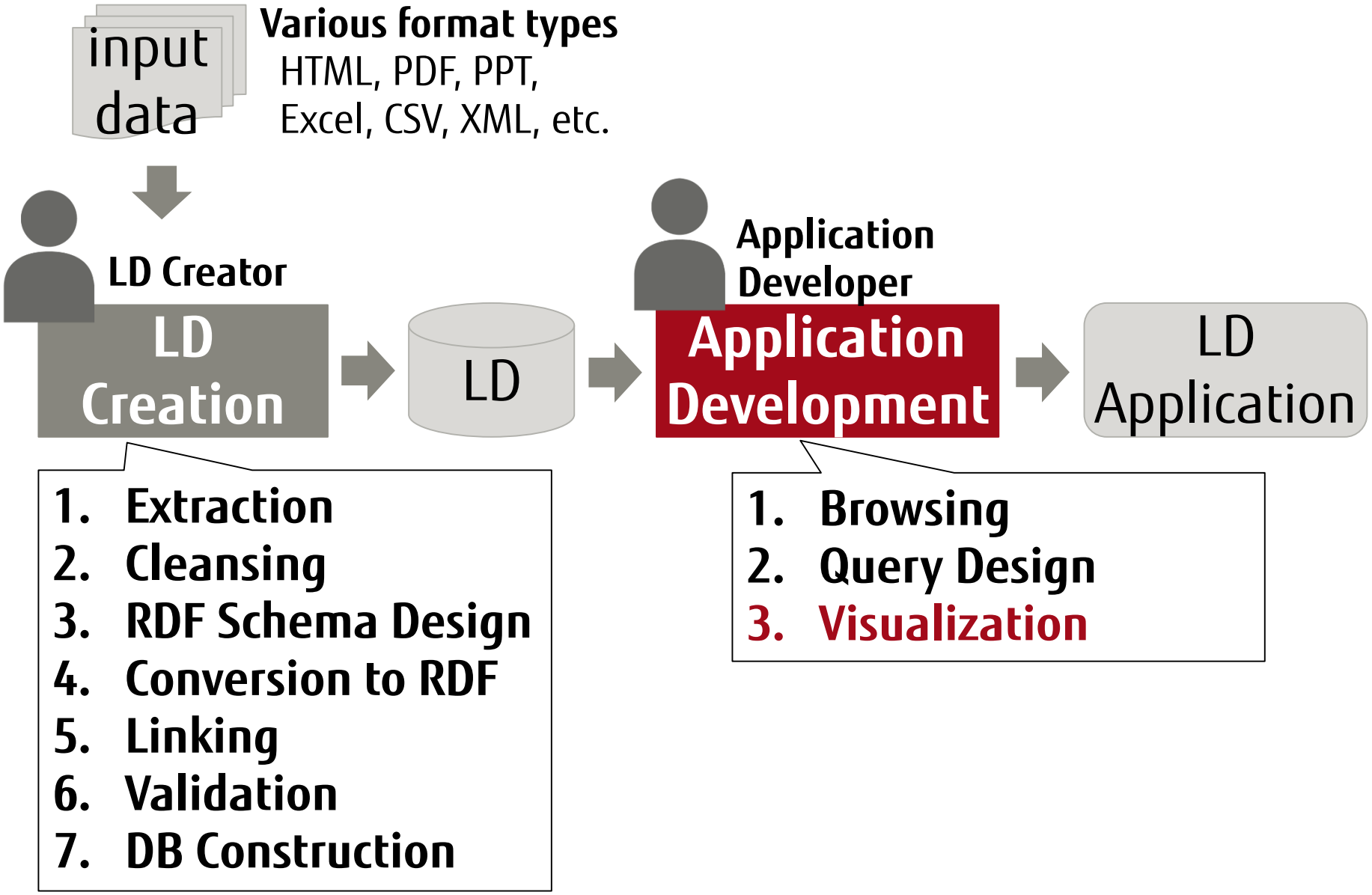
My Favorite Things

FUJITSU

Paragliding near Mt.Fuji



Outline

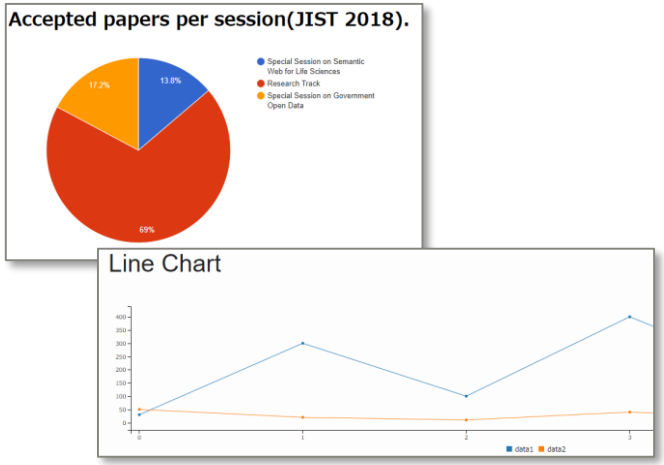


Visualization

FUJITSU

Visualizing the results of SPARQL queries using visualization libraries

SPARQL Result Format → **Visualization** →



- Specialized in SPARQL
  - d3sparql★, Sgvizler
- Standard
  - D3.js★, C3.js
- Wikification
  - TAGME★, DBPedia Spotlight, Illinois Wikifier

- A JavaScript library based on the standard visualization library D3.js.
  - Executing SPARQL query
  - Transforming the resulting JSON for visualization in D3.js layouts
  - Major D3.js layouts
    - Chart: Bar, Pie, Graph: Force, Sankey
    - Tree: Round tree, Tree, map
    - Map, Table

SPARQL query

```
SELECT distinct ?paper
                ?paper_title
                ?author
                ?author_name
WHERE{
        ?paper dc:title ?title;
               dc:creator ?author.
         ?author rdfs:label ?author_name.
       }
```

Searching for JIST 2018 paper titles and authors

Visualization definition

```
function render(json) {
  var config = {
    // for d3sparql.graph()
    "key1": "paper",
    "key2": "authour",
    "label1": "paper_title,
    "label2": "authour_name",
    // for d3sparql.sankey()
    "width":   750,
    "height": 1200,
    "margin":   10,
    "selector": "#result"
  }
  d3sparql.sankey(json, config)
}
```

Specifying the output layout as a Sankey graph

Visualization

Sgvizler

http://mgskjaeveland.github.io/sgvizler/

FUJITSU

- A JavaScript library Rendering the result of SPARQL SELECT queries into charts or HTML elements

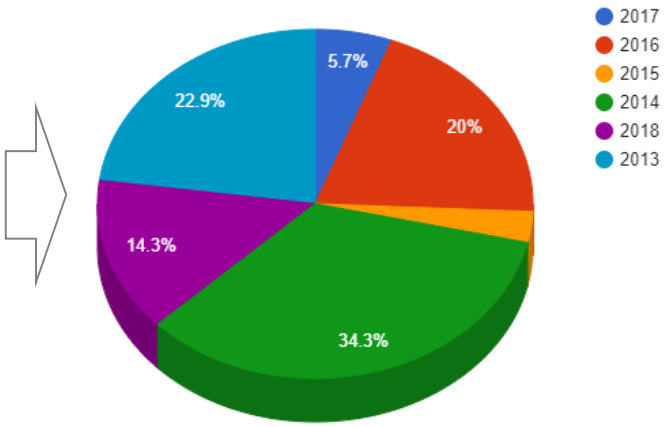
```
<html>
  <head>
    <title>MWE</title>
    <script type="text/javascript" src="//cdnjs.cloudflare.com/ajax/libs/jquery/1.9.0/jque
    <script type="text/javascript" src="http                    pt>
    <script type="text/javascript" src="http                    ler/v/0.6/sgvi
    <script>
      $(document).ready(
          function (){ sgvizler.containerDrawAll(
      );
    </script>
  </head>
  <body>

    <h1>Accepted papers per session(JIST 20    .</h1>
    <p></p>
    <div id="example"
      data-sgvizler-endpoint="http://localhost:3030/jist/query"
      data-sgvizler-query="PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
                           PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
                           SELECT ?subject (count(?paper) as ?c)
                           WHERE {
                           ?paper <http://purl.org/dc/elements/1.1/subject> ?subject .
                           }
                           group by ?subject"
      data-sgvizler-chart="google.visualization.PieChart"
      style="width:800px; height:400px;"></div>
  </body>
</html>
```

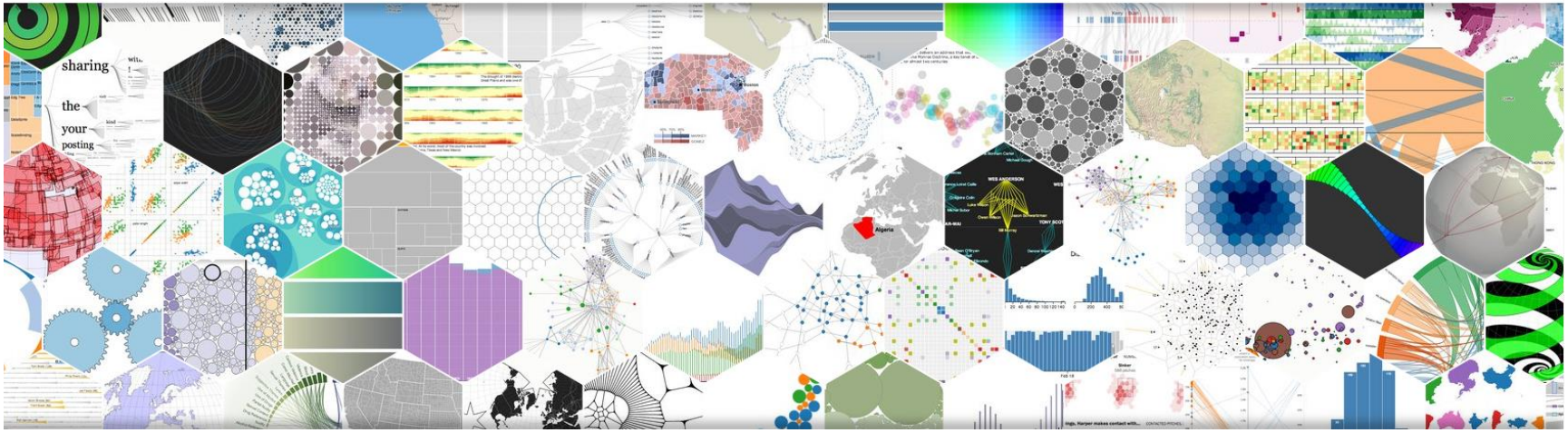
SPARQL endpoint

SPARQL query

Visualization definition:
google.visualization.PicChart



Data-Driven Documents




```
chart = {
  const root = pack(data);

  const svg = d3.select(DOM.svg(width, height))
      .style("width", "100%")
      .style("height", "auto")
      .attr("font-size", 10)
      .attr("font-family", "sans-serif")
      .attr("text-anchor", "middle");

  const leaf = svg.selectAll("g")
    .data(root.leaves())
    .enter().append("g")
      .attr("transform", d => `translate(${d.x + 1},${d.y + 1})`);

  leaf.append("circle")
      .attr("id", d => (d.leafUid = DOM.uid("leaf")).id)
      .attr("r", d => d.r)
      .attr("fill-opacity", 0.7)
      .attr("fill", d => color(d.data.group));

  leaf.append("clipPath")
      .attr("id", d => (d.clipUid = DOM.uid("clip")).id)
    .append("use")
      .attr("xlink:href", d => d.leafUid.href);
```

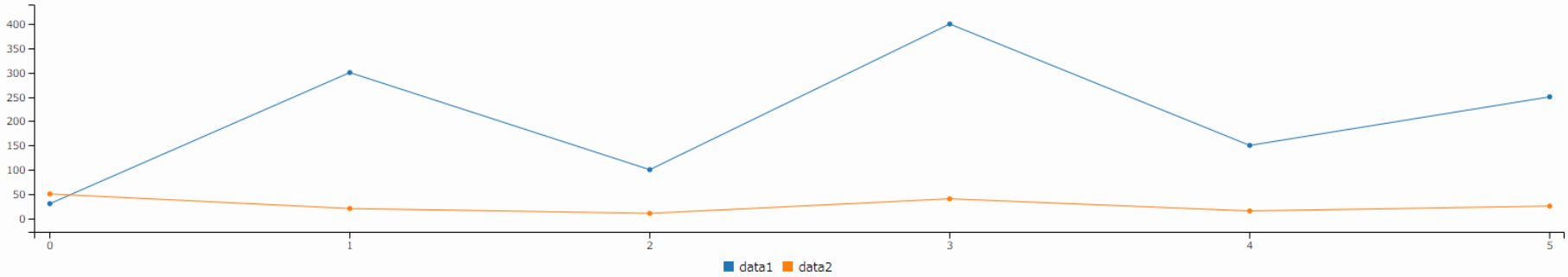
- A JavaScript library for producing dynamic, interactive data visualizations in web browsers
  - Latest version: 5.7.0
  - It helps you bring data to life using HTML, SVG, and CSS

C3.js

https://c3js.org/

- Makes it easy to generate D3-based charts by wrapping the code required to construct the entire chart. We don't need to write D3 code any more.

Line Chart



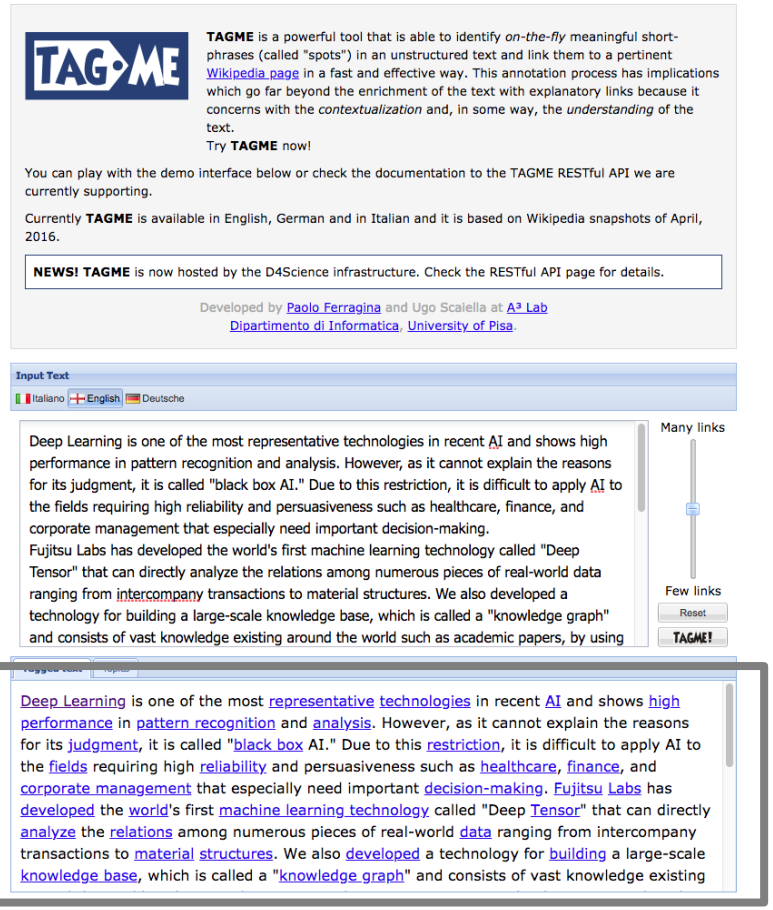
simple_multiple.js

```
var chart = c3.generate({
    data: {
        columns: [
            ['data1', 30, 300, 100, 400, 150, 250],
            ['data2', 50, 20, 10, 40, 15, 25]
        ]
    }
});
```

- Implementation of the Wikification systems
  - available in English, German and in Italian and it is based on Wikipedia snapshots of April, 2016.

Example

In computing, linked data (often capitalized as Linked Data) is a method of publishing structured data so that it can be interlinked and become more useful through semantic queries.[1] It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers. Part of the vision of linked data is for the internet to become a global database.



TAGME (cont.)

Visualization (5/7)

https://tagme.d4science.org/tagme/

FUJITSU

- Ex. Annotated titles

  - Ontology-based Semantic Representation of Spatial Configuration: Shape Grammar for Caravanserai Architectural Heritage
  - DeFind: A Protege Plugin for Computing Concept Definitions in EL Ontologies

DBpedia Spotlight

Visualization (6/7)

https://www.dbpedia-spotlight.org/demo/index.html

FUJITSU

- A tool for automatically annotating mentions of DBpedia resources in text

Example

In computing, linked data (often capitalized as Linked Data) is a method of publishing structured data so that it can be interlinked and become more useful through semantic queries.[1] It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers. Part of the vision of linked data is for the internet to become a global database.

Illinois Wikifier

Visualization (7/7)

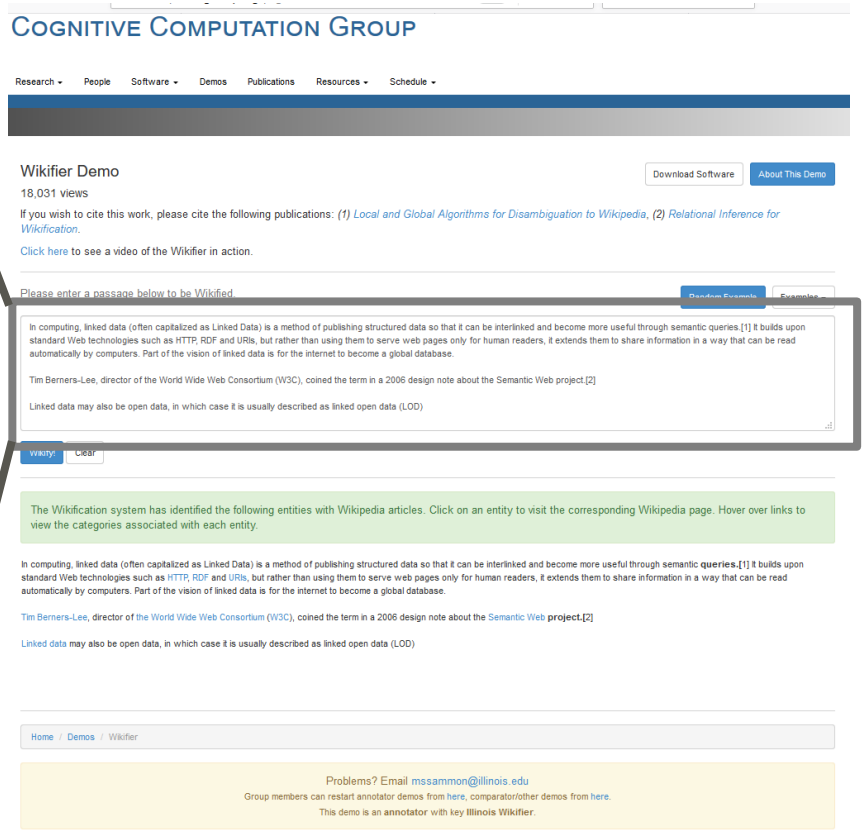
https://cogcomp.org/page/download_view/Wikifier

FUJITSU

- also makes links with Wikipedia

Example

In computing, linked data (often capitalized as Linked Data) is a method of publishing structured data so that it can be interlinked and become more useful through semantic **queries.[**1] It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers. Part of the vision of linked data is for the internet to become a global database.



My Favorite Things

FUJITSU

Liver Sashimi

FUJITSU

Final Remarks

- Let's Review Through The Sample Application

JIST LD Application

FUJITSU

Trends of International Conferenc

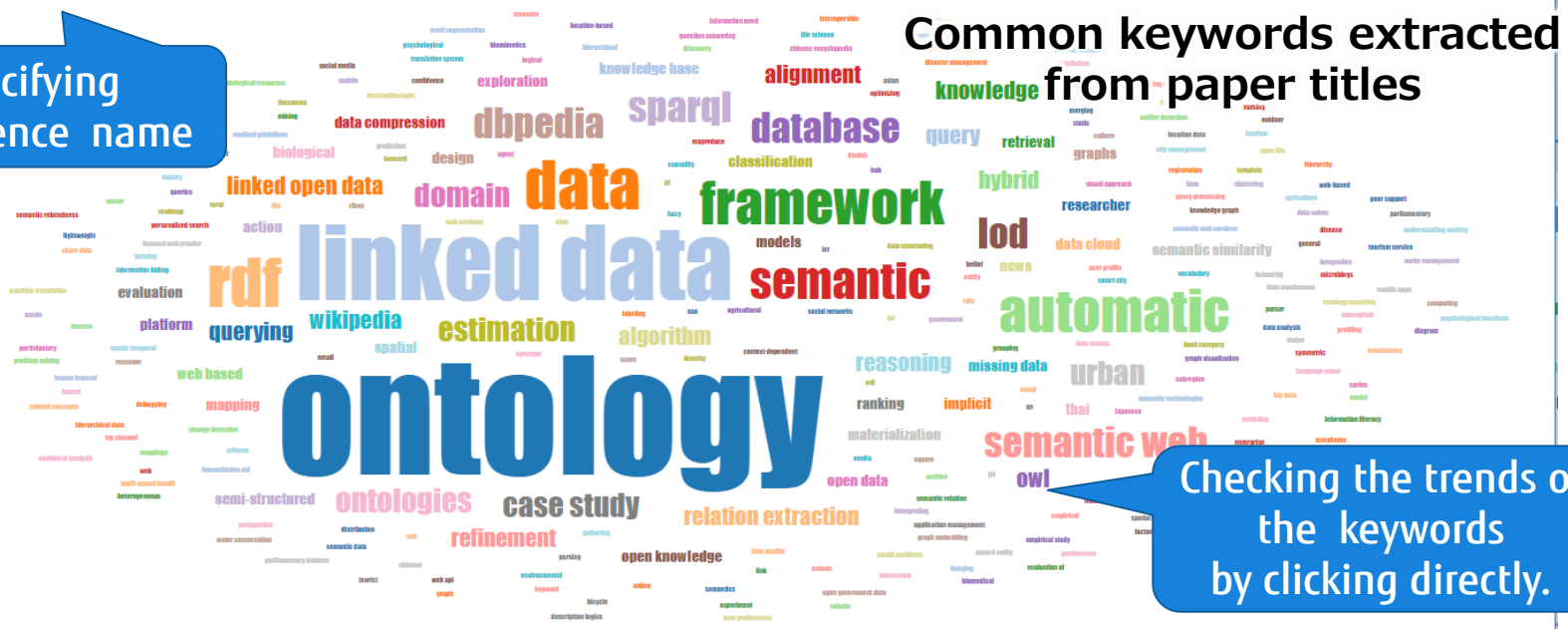
localhost

COMMON KEYWORDS EXTRACTED FROM PAPERS TITLES

JIST Search

Specifying conference name

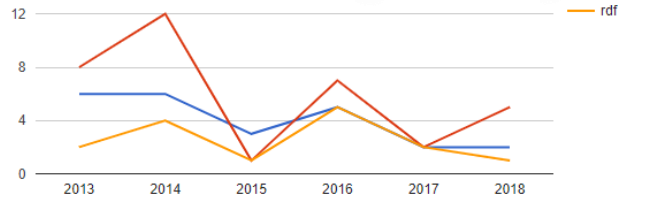
Common keywords extracted from paper titles



Checking the trends of the keywords by clicking directly.

TRENDS OF CLICKING KEYWORDS

Trend of users clicking the keyword



PAPER TITLES CONTIAN A SELECTED KEYWORD

Paper titles containing the keyword

- Building Wiki[illegible] Ontology with More S[illegible]ured Information Resources.
- Ontology Con[illegible]e[illegible]al
- SQuaRE: A Vi[illegible]ol[illegible]s.
- Development of Faceted [illegible]nd Synonym Search Support for the Ontology Application Management Framework.
- Development of the Belief Culture Ontology and Its Application: Case Study of the GreaterMekong Subregion.
- Automatic and Dynamic Book Category Assignment Using Concept-Based Book Ontology.
- Ontology Based Suggestion Distribution System.
- Constructing City Ontology from Expert for Smart City Management.
- Ontology Construction Support for Specialized Books.
- A Graph-Based Approach to Ontology Debugging in DL-Lite.
- Utilizing Weighted Ontology Mappings on Federated SPARQL Querying.
- A Framework of Automatic Alignment of Concept in Ontology with Confidence Score based on Inner Concept Information.
- MAPSOM: User Involvement in Ontology Matching.
- Designing of Ontology for Domain Vocabulary on Agriculture Activity Ontology (AAO) and a Lesson Learned.
- An Information Literacy Ontology and Its Use for Guidance Plan Design - An Example on Problem Solving.
- Ontology Schema-specific Rule Materialization.
- Ontology Refinement System for Improving Consistency of Classification among Brother Concepts.
- Federating Heterogeneous Biological Resources on the Web: A Case Study on TRP Channel Ontology Construction

JIST LD Creation

FUJITSU

Extraction ★Beautiful Soap

Cleasing ★Open Refine

RDF Schema Design ★Linked Open Vocabraries ★Google Dataset Search ★Data Sniffier

Conversion to RDF ★Open Refine +RDF Extension

Linking ★Silk

Validation ★rapper ★arq ★SHACL

DB Construction ★Virtuoso

RDF(.ttl)

```
jist:xxx
  dblp:identifier “id_x” ;
  dblp:authorBy dblp:author_x ;
  dblp:tilte “title_x” .
```

SPARQL endpoint

DB

JIST LD Application Development

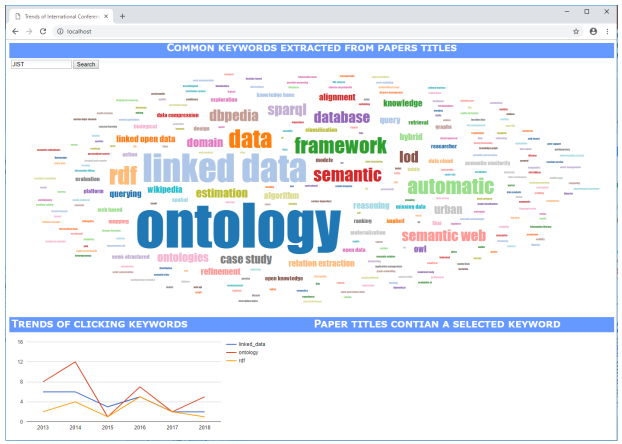
FUJITSU

RDF(.ttl)

```
jist:xxx
  dblp:identifier “id_x” ;
  dblp:authorBy dblp:author_x ;
  dblp:tilte “title_x” .
```

SPARQL endpoint

DB



- **Browsing**
  - ★Pubby
  - ★GraphViz
  - ★Visual RDF
- **Query Design**
  - ★rsparql
- **Visualization**
  - ★d3sparql
  - ★d3.js
  - ★Tagme

JIST LD Application : Detail

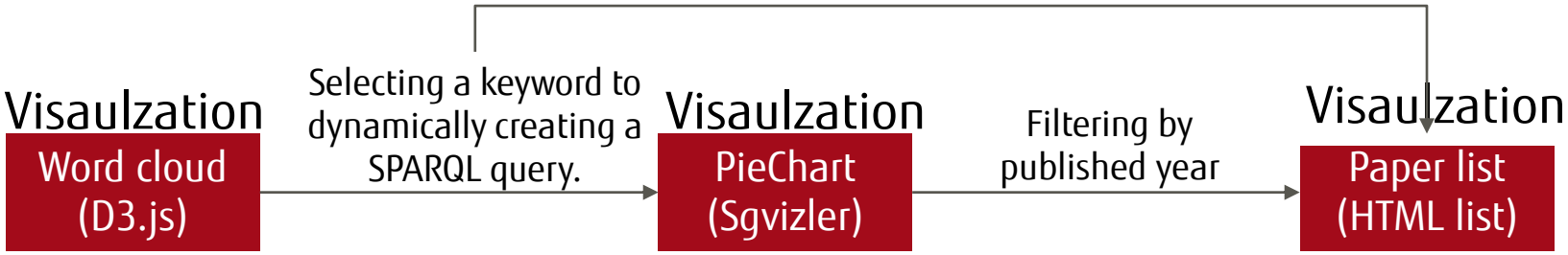
FUJITSU

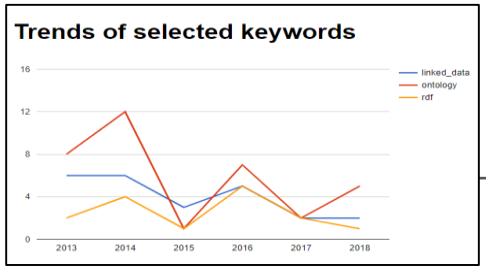
TAGME + D3.js

JIST LD Application : Detail (cont.)

FUJITSU

- D3.js + Sgvizler



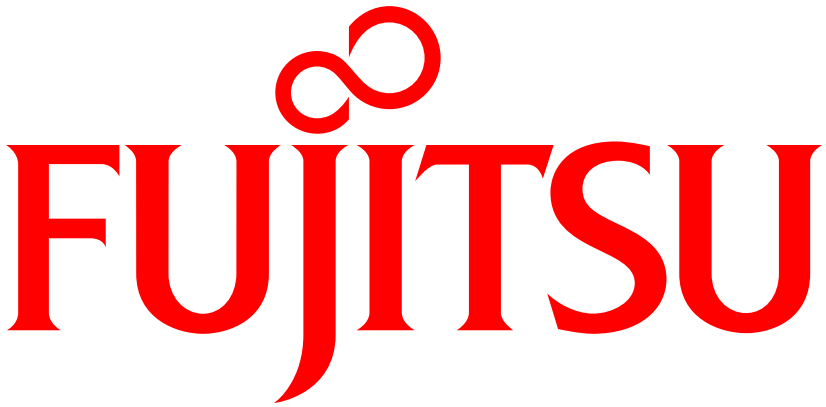


- Building Wikipedia Ontology with More Semi-structured Information Resources.
- Ontology Construction and Schema Merging Using an Application-Independent Ontology for H
- SQuaRE: A Visual Approach for Ontology-Based Data Access.
- Development of Faceted and Synonym Search Support for the Ontology Application Managen
- Development of the Belief Culture Ontology and Its Application: Case Study of the GreaterMel
- Automatic and Dynamic Book Category Assignment Using Concept-Based Book Ontology.
- Ontology Based Suggestion Distribution System.
- Constructing City Ontology from Expert for Smart City Management.
- Ontology Construction Support for Specialized Books.
- A Graph-Based Approach to Ontology Debugging in DL-Lite.

Introduction : Schedule

FUJITSU

Morning	Introduction to Linked Data tools	9:00-10:00 (60 mins)	Presentation
		10:00-10:10 (10 mins)	Break
		10:10-10:50 (40 mins)	Presentation
		10:50-11:10 (20 mins)	Use Case Studies
Afternoon	Hands-on	13:00-13:10 (10 mins)	Introduction
		13:10-14:55 (95 mins)	Trial 1: Let's visualize corporate data Trial 2: Let's develop a financial analysis application
		14:55-15:00 (5 mins)	Closing



shaping tomorrow with you

Glossary

FUJITSU

- cURLA: command line Open Source/Free Software client that can transfer data, including machine readable RDF, from or to a server using one of its many supported protocols.
- DBLP:DBLP is a computer science bibliography website. Starting in 1993 at the University of Trier, Germany, it grew from a small collection of HTML files[1] and became an organization hosting a database and logic programming bibliography site. DBLP listed more than 3.66 million journal articles, conference papers, and other publications on computer science in July 2016, up from about 14,000 in 1995.[2] All important journals on computer science are tracked. Proceedings papers of many conferences are also tracked. It is mirrored at three sites across the Internet
- Dbpedia: DBpedia is a community effort to extract structured information from Wikipedia and make it available on the Web. DBpedia is often depicted as a hub for the Data Cloud. An RDF representation of the metadata held in Wikipedia and made available for SPARQL query on the World Wide Web.
- Dublin Core Metadata Element Set : Dublin Core Metadata Element Set refers to a vocabulary of fifteen properties for use in resource descriptions, such as may be found in a library card catalog (creator, publisher, etc). The Dublin Core Metadata Element Set, also known as "DC Elements", is the most commonly used vocabulary for Linked Data applications. See also Dublin Core Element Set, Version 1.1 Specification. [DCMI]
- Dublin Core Metadata Terms : A vocabulary of bibliographic terms used to describe both physical publications and those on the Web. An extended set of terms beyond those basic terms found in the Dublin Core Metadata Element Set. See also Dublin Core Metadata Terms
- Entity : In the sense of an entity-attribute-value model, an entity is synonymous with the Subject of an RDF Triple.
- FOAF (Friend of a Friend) A Semantic Web vocabulary describing people and their relationships for use in resource descriptions.
- GitHub: **GitHub Inc.** is a web-based hosting service for version control using Git. It is mostly used for computer code. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for every project

Glossary (cont.)

FUJITSU

- **International Semantic Web Conference**(ISWC): The International Semantic Web Conference (ISWC) is a series of academic conferences and the premier international forum, for the Semantic Web, Linked Data and Knowledge Graph Community.[1][2][3] Here, scientists, industry specialists, and practitioners meet to discuss the future of practical, scalable, user-friendly, and game changing solutions.[4] Its proceedings are published in the Lecture Notes in Computer Science by Springer-Verlag.
- JSON : JavaScript Object Notation (JSON) is syntax for storing and exchanging text based information. JSON has proven to be a highly useful and popular object serialization and messaging format for the Web. See also: the application/json Media Type for JavaScript Object Notation (JSON) [RFC4627].
- Linked Data (LD): A pattern for hyperlinking machine-readable data sets to each other using Semantic Web techniques, especially via the use of RDF and URIs. Enables distributed SPARQL queries of the data sets and a browsing or discovery approach to finding information (as compared to a search strategy). Linked Data is intended for access by both humans and machines. Linked Data uses the RDF family of standards for data interchange (e.g., RDF/XML, RDFa, Turtle) and query (SPARQL). If Linked Data is published on the public Web, it is generally called *Linked Open Data*. See also [Linked Data Principles].
- **Linked Open Data**(LOD): Linked Data published on the public Web and licensed under one of several open licenses permitting reuse. Publishing Linked Open Data enables distributed SPARQL queries of the data sets and a "browsing" or "discovery" approach to finding information, as compared to a search strategy. See also: "Linked Data: Structured Data on the Web" [LD-FOR-DEVELOPERS] and "Linked Data: Evolving the Web into a Global Data Space" [HOWTO-LODP]
- Ontology: A formal model that allows knowledge to be represented for a specific domain. An ontology describes the types of things that exist (classes), the relationships between them (properties) and the logical ways those classes and properties can be used together (axioms).

Glossary (cont.)

FUJITSU

- **Web Ontology Language (**OWL): The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies. Ontologies are a formal way to describe taxonomies and classification networks, essentially defining the structure of knowledge for various domains: the nouns representing classes of objects and the verbs representing relations between the objects. Ontologies resemble class hierarchies in object-oriented programming but there are several critical differences. Class hierarchies are meant to represent structures used in source code that evolve fairly slowly (typically monthly revisions) whereas ontologies are meant to represent information on the Internet and are expected to be evolving almost constantly. Similarly, ontologies are typically far more flexible as they are meant to represent information on the Internet coming from all sorts of heterogeneous data sources. Class hierarchies on the other hand are meant to be fairly static and rely on far less diverse and more structured sources of data such as corporate databases
- Python: Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.[27] In July 2018, Van Rossum stepped down as the leader in the language community after 30 years
- **Resource Description Framework (RDF): A family of international standards for data interchange on the Web produced by W3C. Resource Description Framework (RDF) is based on the idea of identifying things using Web identifiers or HTTP URIs, and describing resources in terms of simple properties and property values. See also [RDF 1.1 Concepts and Abstract Syntax**
- **RDF Schema (**RDFS):The simplest RDF vocabulary description language. It provides much less descriptive capability than the Simple Knowledge Organization System (SKOS) or the Web Ontology Language (OWL). A standard of the W3C [RDFS]
- SPARQL: SPARQL Protocol and RDF Query Language (SPARQL) defines a query language for RDF data, analogous to the Structured Query Language (SQL) for relational databases. A family of standards of the World Wide Web Consortium. See also SPARQL 1.1 Overview [SPARQL-11].

Glossary (cont.)

FUJITSU

- SPARQL endpoint: A service that accepts SPARQL queries and returns answers to them as SPARQL result sets. It is a best practice for datasets providers to give the URL of their SPARQL endpoint to allow access to their data programmatically or through a Web interface. A list of some endpoints status is available at http://labs.mondeca.com/sparqlEndpointsStatus/
- **Semantic Web Rule Language(**SWRL) The **Semantic Web Rule Language** (**SWRL**) is a proposed language for the Semantic Web that can be used to express rules as well as logic, combining OWL DL or OWL Lite with a subset of the Rule Markup Language (itself a subset of Datalog).
- YAML: YAML (YAML Ain't Markup Language) is a human-readable data serialization language. It is commonly used for configuration files, but could be used in many applications where data is being stored (e.g. debugging output) or transmitted (e.g. document headers). YAML targets many of the same communications applications as XML but has a minimal syntax which intentionally breaks compatibility with SGML .[1] It uses both Python-style indentation to indicate nesting, and a more compact format that uses [] for lists and {} for maps[1] making YAML 1.2 a superset of JSON.[2]

References

- Wikipedia: https://en.Wikipedia.org/
- **Linked Data Glossary**: https://www.w3.org/TR/ld-glossary/