Towards a Semantic Toolbox for Reproducible Knowledge Graph Generation in the Biodiversity Domain – How to Make the Most out of Biodiversity Data

Samira Babalou

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

David Schellenberger Costa

UNIVERSITÄT LEIPZIG

Jens Kattge

Max Planck Institute for Biogeochemistry

Christine Römermann

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Birgitta König-Ries

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

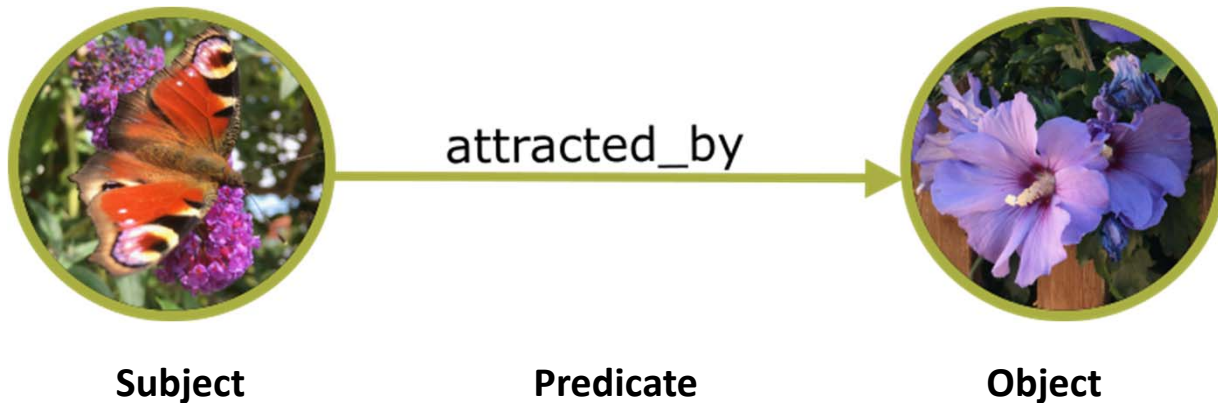**CS4BioDiversity Workshop at INFORMATIK2021**

27.09.2021

#INFORMATIK 2021

Outline
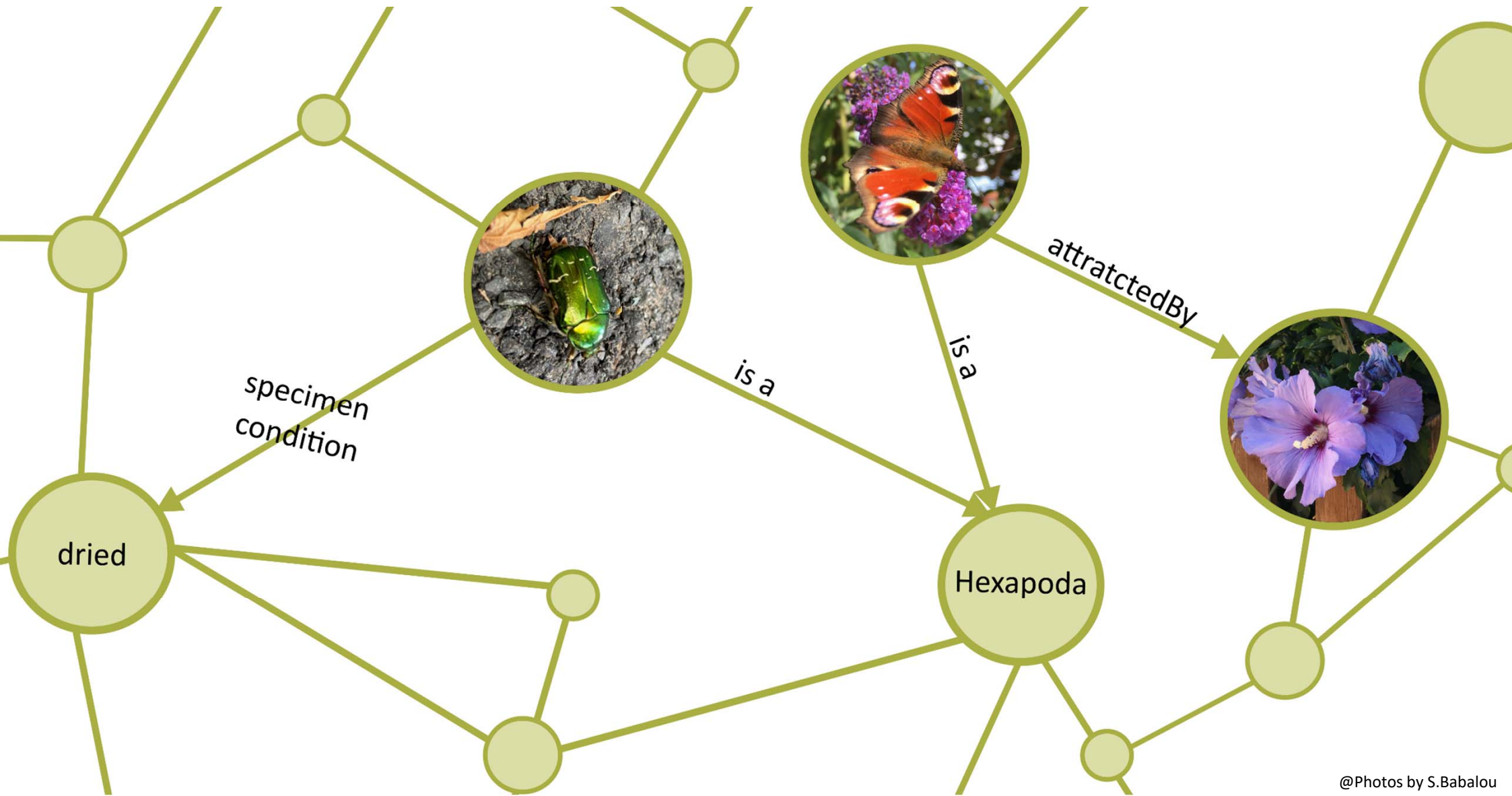
- Background knowledge
- State of the art
- Proposed workflow
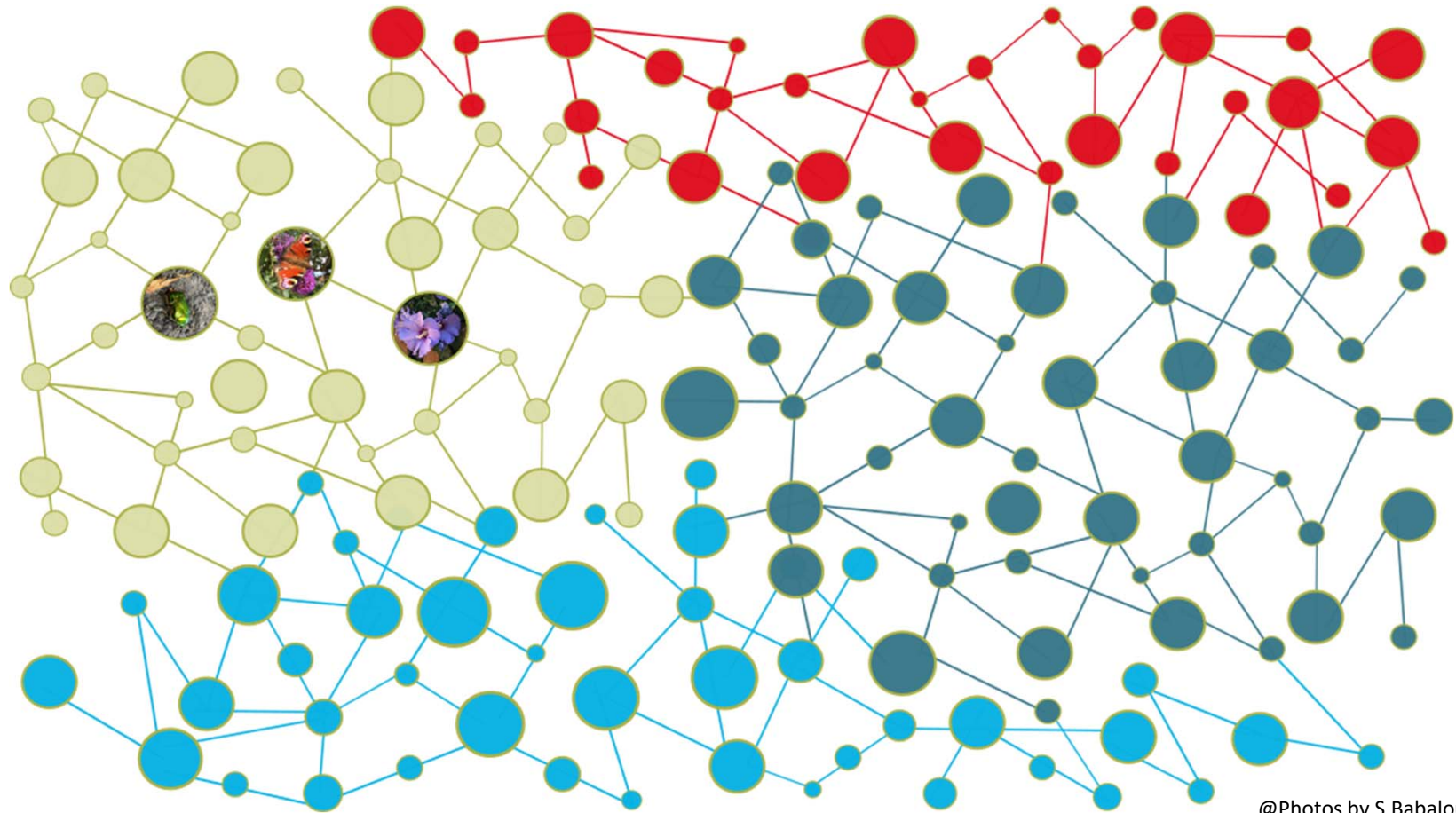- Outlook

Unique Knowledge Representation by RDF Triples

Subject   **Predicate**   **Object**

attracted_by

RDF (Resource Description Framework)

```
<Butterfly>   <attractedBy>   <Flower>
<Flower>      <support>       <Pollinators>
<Butterfly>   <is_a>          <Hexapoda>
...
```

@Photos by S.Babalou

attratctedBy

specimen condition

is a

is a

dried

Hexapoda

@Photos by S.Babalou

Knowledge Graph (KG)

[Nickel et al., 2015]

@Photos by S.Babalou

Benefits of KGs

① Interlink

② Query

🎤 Show me all species in botanical garden Jena? 🔍

③ Inference

@Photos by S.Babalou

Leveraging the KGs in the biodiversity domain: iKNOW project

within the German Centre for Integrative Biodiversity Research (iDiv)

https://fusion.cs.uni-jena.de/fusion/projects/iknow/
https://www.idiv.de/en/index.html

Building KGs from Tabular Data

Table 1

Table 2

Unique Knowledge Representation by RDF Triples

attracted_by

Subject   **Predicate**   **Object**

<Butterfly>   <attractedBy>   <Flower>
<Flower>   <support>   <Pollinators>
<Butterfly>   <is_a>   <Hexapoda>
...

@Photos by S.Babalou
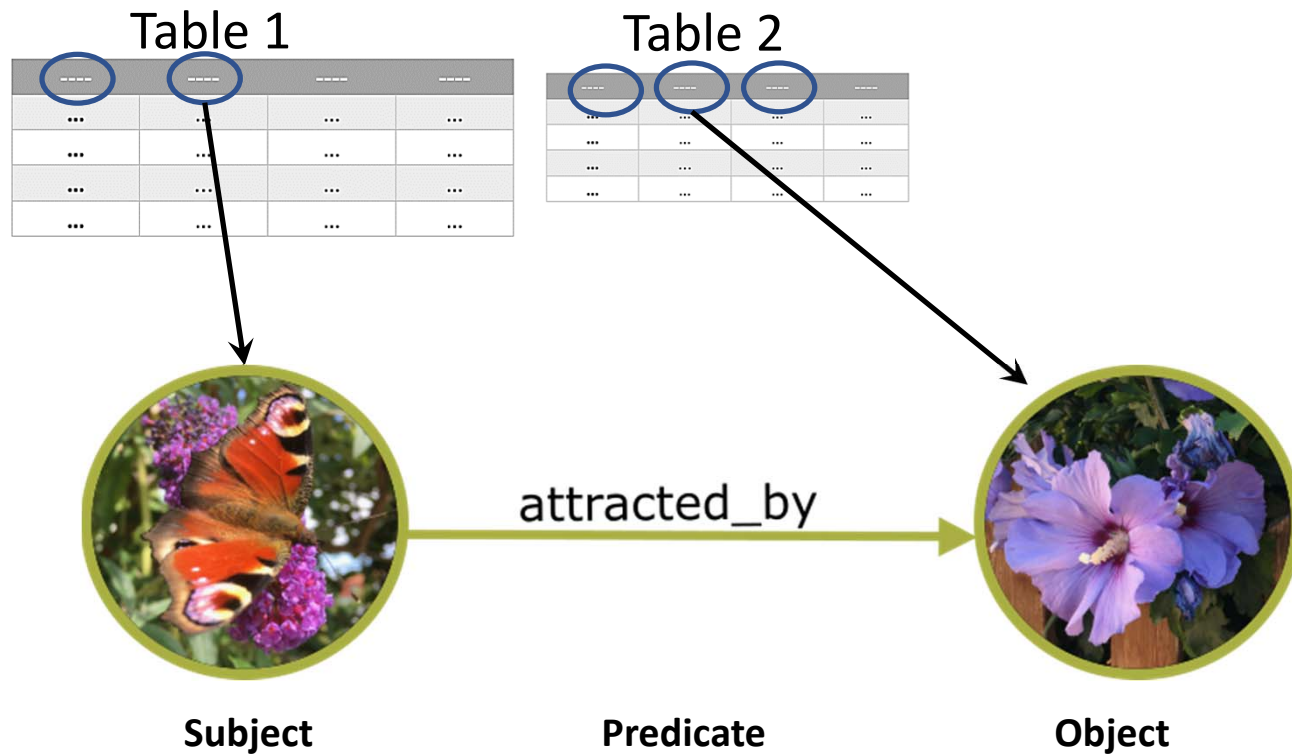
Biodiversity Datasources

TRY
[Kattge et al., 2011]

Is the main plant trait database used by the research community worldwide.

sPlot
[Bruelheide et al., 2019]

Explores global patterns of vegetation plots worldwide.

PhenObs
[Nordt et al., 2021]

Discovers the impacts of climate change on the phenology of herbaceous plant species.

PlantHub

Aims to integrate these datasets in national and international developments including NFDI4BioDiversity

https://www.lw.uni-leipzig.de/en/institute-for-biology/abteilungen/systematic-botany-and-functional-biodiversity/research/big-data/

State of the Art

- Biodiversity KGs
  - Ozymandias [R. Page, 2019]
  - OpenBiodiv [Penev et al., 2019]

- Dynamic cross-domain KGs
  - DBPedia [Auer et al., 2007] and WikiData [Vrandečić et al., 2014]

- Efforts of building dynamic KGs
  - [Ji et al., 2015, He et al., 2017, Chang et al., 2017]

Proposed Approach

Provenance tracking

Pre-Processing

Data cleaning & normalization

RDFization

Triple Generation of tabular data + cross linking

Graph Database Repository

ACCESSIBILITY

Simple Search

SPARQL Search (semantic)
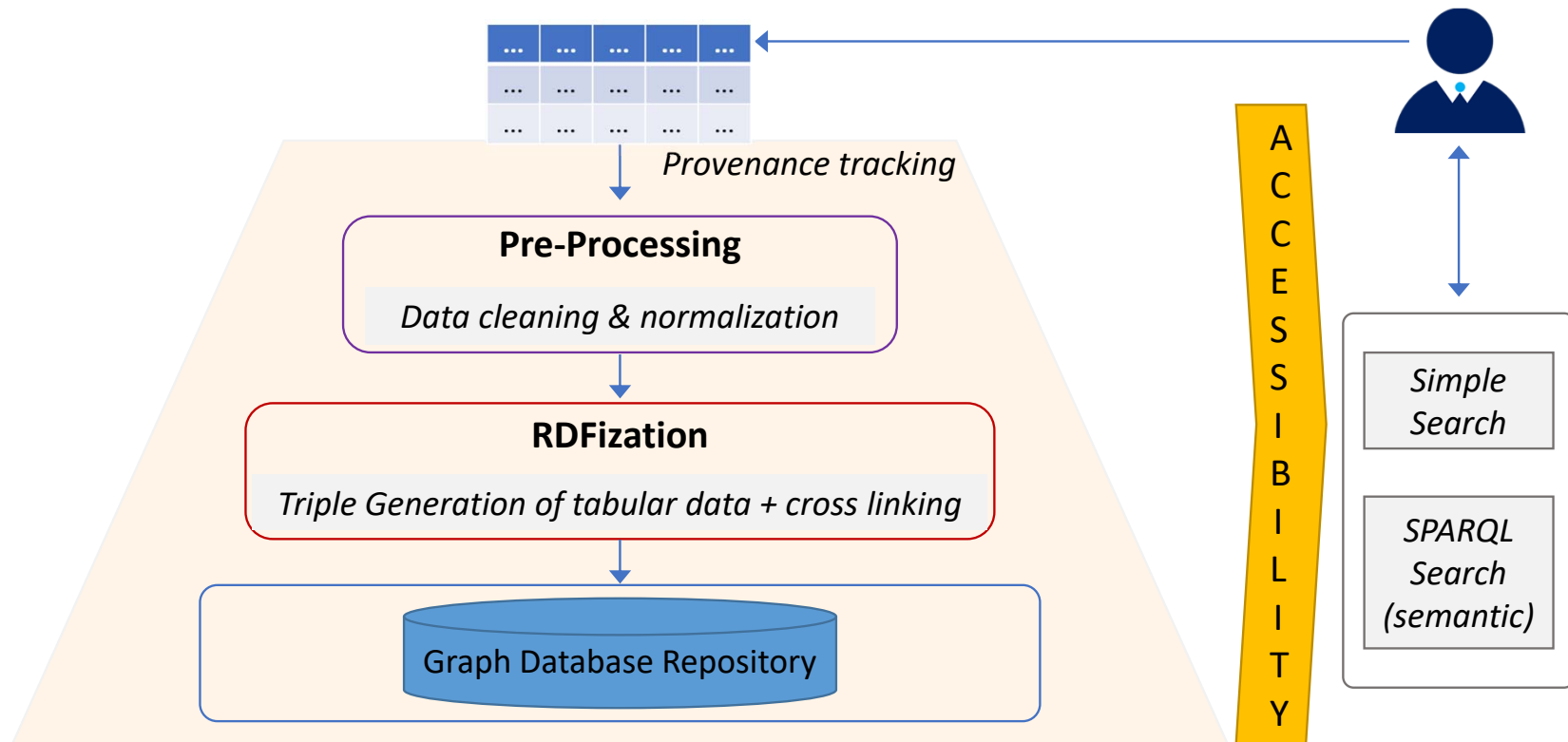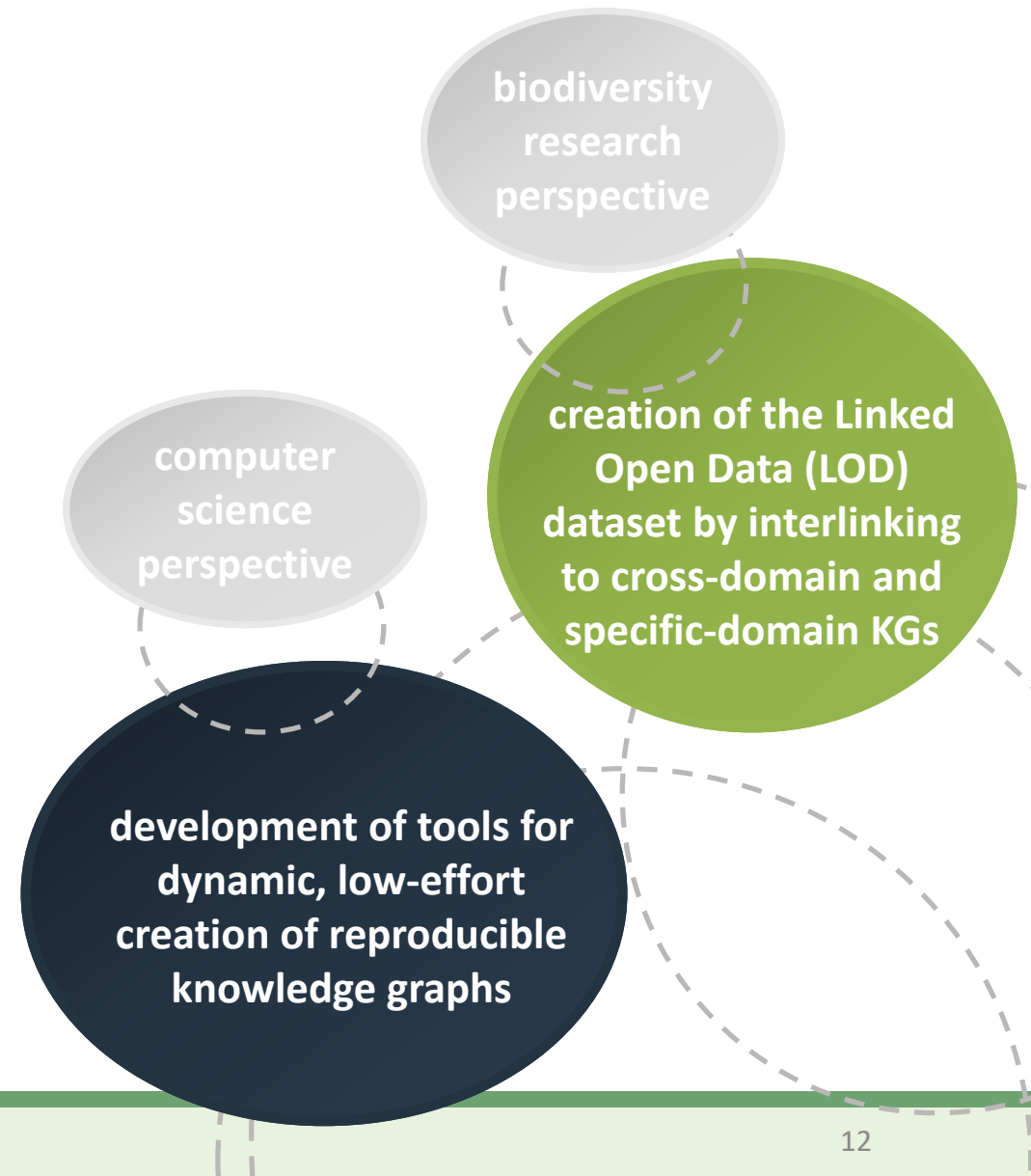
JenTab [Abdelmageed &Schindler, 2020]

Outlook

- iKNOW

  is a planned semantic-based toolbox for the dynamic building Knowledge Graph in the biodiversity domain

- Provenance tracking & reproducibility functionality

  establishment of open science practices in the biodiversity domain

biodiversity research perspective

computer science perspective

creation of the Linked Open Data (LOD) dataset by interlinking to cross-domain and specific-domain KGs

development of tools for dynamic, low-effort creation of reproducible knowledge graphs

Special Thanks to

iDiv

German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig

THANK YOU

✉ Samira.Babalou@uni-jena.de

🐦 @SamiraBabalou

in Samira-Babalou

Ⓢ Samira.Babalou

iD 0000-0002-4203-1329

How to Make the Most out of Biodiversity Data