

Alohomora!

Facilitating Personal Data Access

Through Automated Graph Extraction

and Integration

Nicola Leschke¹, Karl Wolf², Frank Pallas¹

¹Privacy Engineering and Policy-Aligned Systems (PEPSys) Group, University of Salzburg, Austria

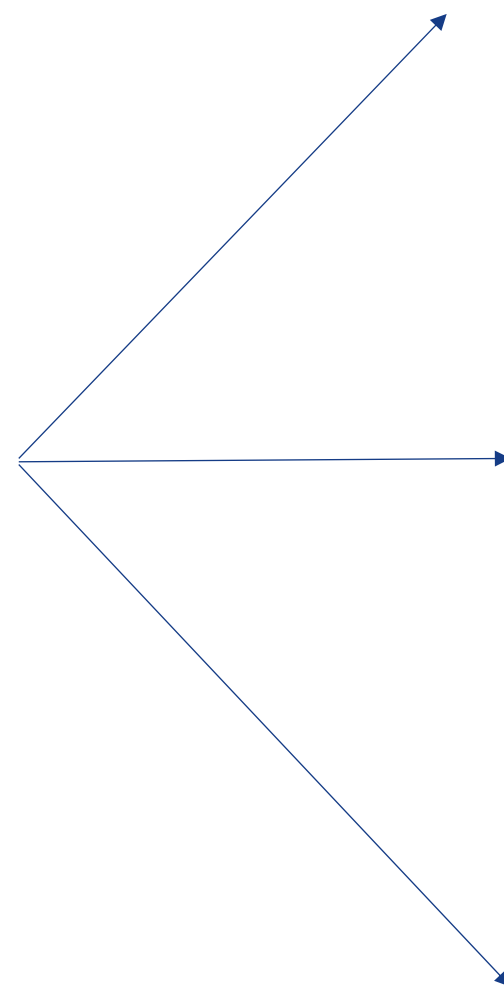
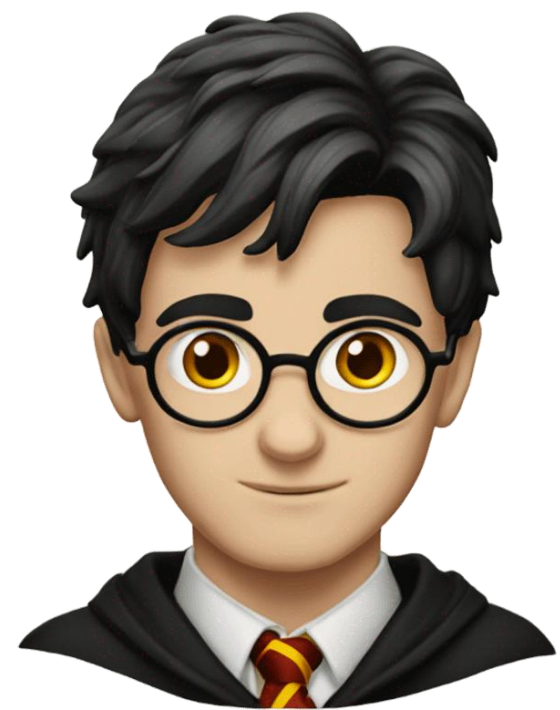
²Information Systems Engineering (ISE) Group, TU Berlin, Germany

Alohomora!

Unlocking the Prophecy (=Personal Data)



Unlocking Personal Data: Beyond the Prophecy



What does the ministry know about me?

→ General discovery of personal data

What is in the prophecy that is stored in the ministry?

→ Specific piece of information

The death eaters took over the ministry, Hogwarts and the magical world. What could they learn when merging the data from different sources?

→ Merging data sources

Without Magic: Personal Data Access



Harry can request a copy of personal data under Art. 15 or Art. 20 GDPR.

Art. 15: Personal Data Access

- fundamental right in privacy legislation
- informed privacy decisions
- prerequisite for exercising other rights

Right to:

- receive a confirmation if personal data processed
- if so, provide access to personal data
- including a copy of personal data

Art. 20: Data Portability

introduced by the GDPR:

- „unlock“ personal data
- simplify transition to privacy-friendly companys

Right to:

- retrieve a **machine-readable** copy of provided data
- transfer that data

DSARs: Data Access and Data Portability

Rights to get a copy of personal data

Art. 15 (3) right to access

→ all personal data

→ electronic format

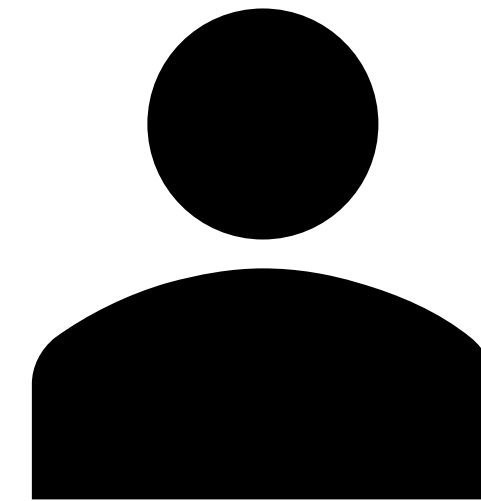
Art. 20 (1) right to data portability

→ personal data provided

→ machine-readable copy

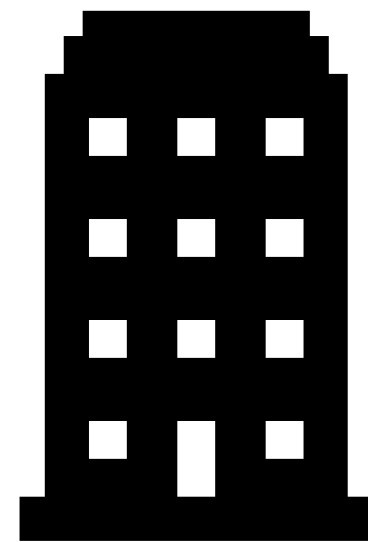
main focus: Art. 15 (3) GDPR
however, it is transferable to Art. 20 (1)
and other legal obligations (CCPA, DMA,
Data Act,...).

Who has ever requested a copy of personal data?



data subject

Who has ever answered a DSAR?



controller

Background: The Response



How do DSAR responses look like?

Example: Facebook

In practice, resulting copies of DSARs according to Art. 15 and Art. 20 are similar.

(Data) Subject Access Request Package
(SARP)

The Challenge: From SARP to Information



Are SARPs sufficient for Harry to understand what data is there?

SARPs are Underused

Data subjects struggle to use SARPs

lack of...

- understandability^[5]
- data integration^[4]
- data reuse^[2]

due to...

- syntactic,
 - structural,
 - semantic
- heterogeneity^[6]

whish for...

- unified data representation^[1]
- data visualization^[3]
- query and filtering^[5]

¹ Borem, A., Pan, E., Obielodan, O., Roubinowitz, A., Dovichi, L., Mazurek, M.L., Ur, B.: Data subjects' reactions to exercising their right of access. In: Proceedings of the 33rd USENIX Security Symposium (2024)

² Syrmoudis, E. et al.: Unlocking personal data from online services: user studies on data export experiences and data transfer scenarios. Human-Computer Interaction pp. 1–25 (2024).

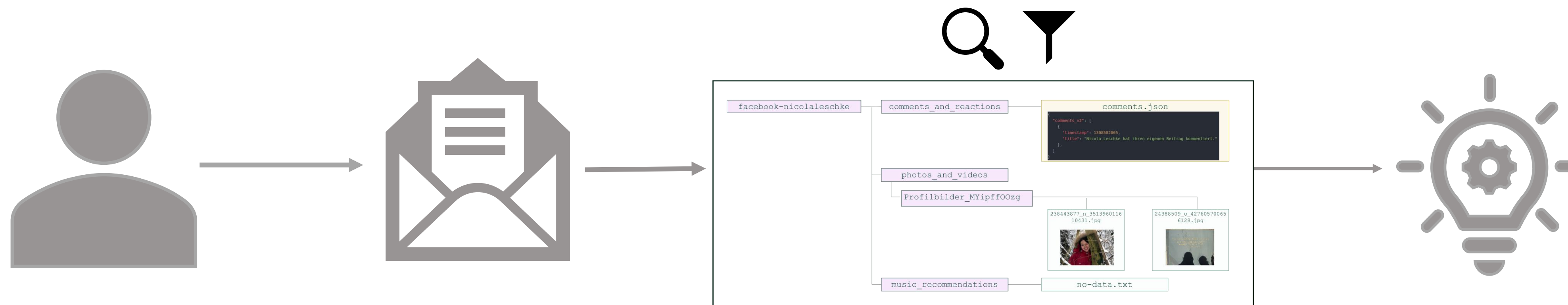
³ Tolsdorf, J., Fischer, M., Lo Iacono, L.: A Case Study on the Implementation of the Right of Access in Privacy Dashboards. In: Privacy Technologies and Policy. pp. 23–46. Springer International Publishing, Cham (2021)

⁴ Turner, S., Galindo Quintero, J., Turner, S., Lis, J., Tanczer, L.M.: The exercisability of the right to data portability in the emerging internet of things (IoT) environment. New Media & Society 23(10), 2861–2881 (2021)

⁵ Veys et al: Pursuing usable and useful data downloads under GDPR/CCPA access rights via co-design (2021)

⁶ Sheth: "Changing focus on interoperability in information systems: From system, syntax, structure to semantics," in Interoperating geographic information systems, Springer, (1999),

Research Gap: Data Representation



response: SARP

Subject Access Request Package

Objective: Find a **unified** machine-readable **representation** of SARPs.

Preliminaries

Structural Description

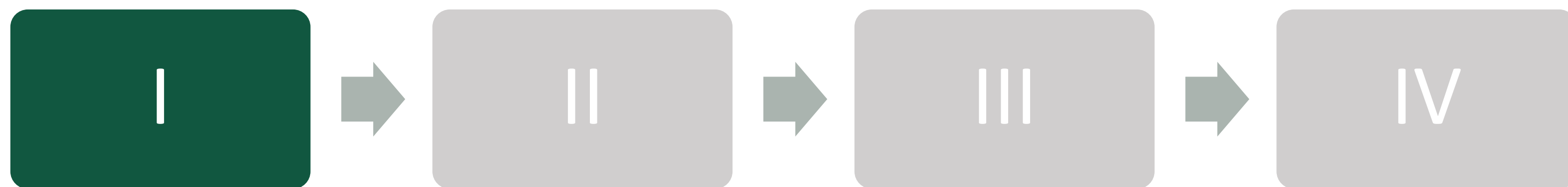
- SARP = package:
 - hierarchical directory
 - (semi-structured) files
- hierarchical structures within files:
 - lists
 - maps
 - tables = maps x lists

Representing SARPs

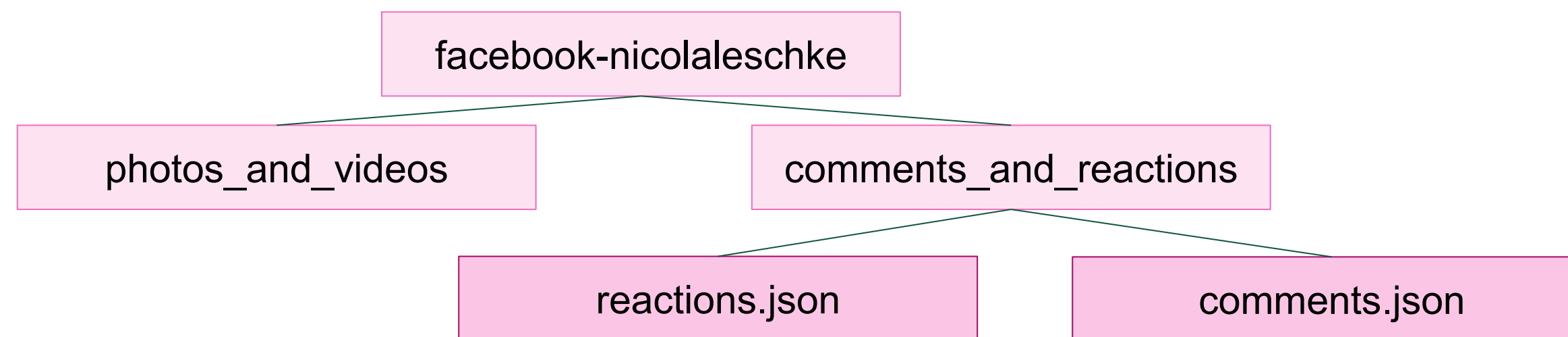
- (R1) Preservation of controller-specific structures
- (R2) Adaptive to unknown structures
- (R3) Many to Many Relations
- (R4) Easily query- and filterable
- (R5) Little to no effort for data subjects

Approach

Iterative construction of a unified representation:



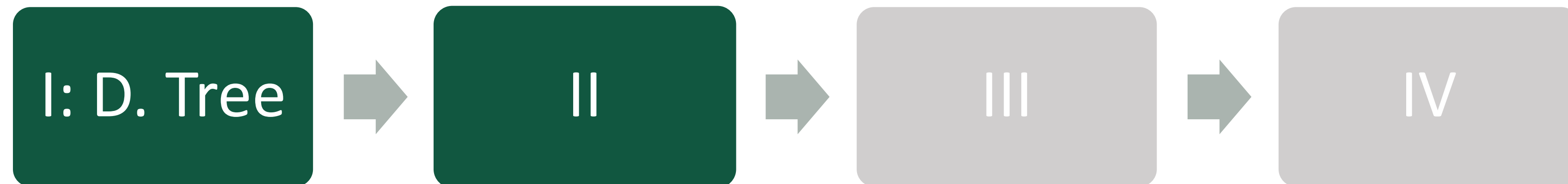
I: Tree-based Directory (Spanning Tree)



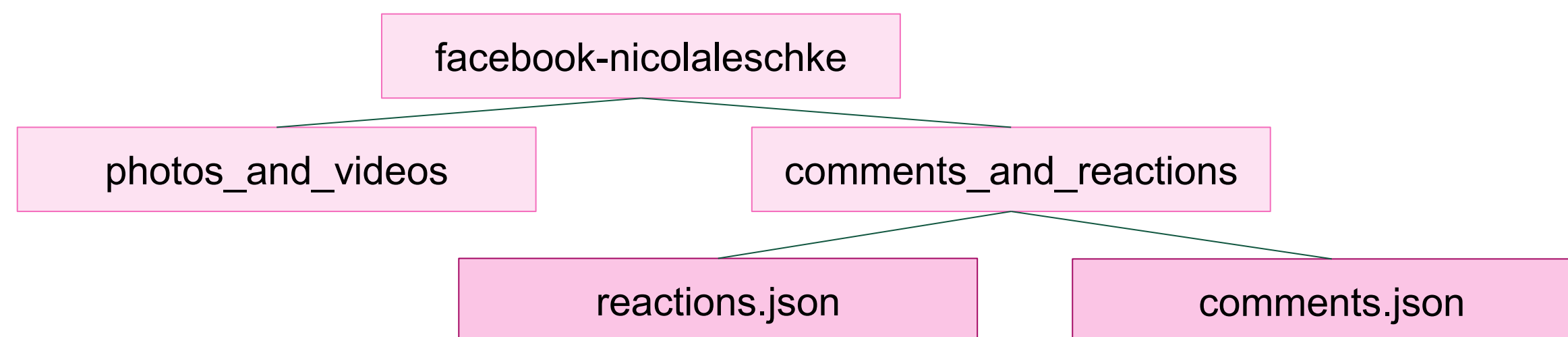
Implementation	
node	node type
folder	empty, folder
file	(semi-)structured, text, media, other

Approach

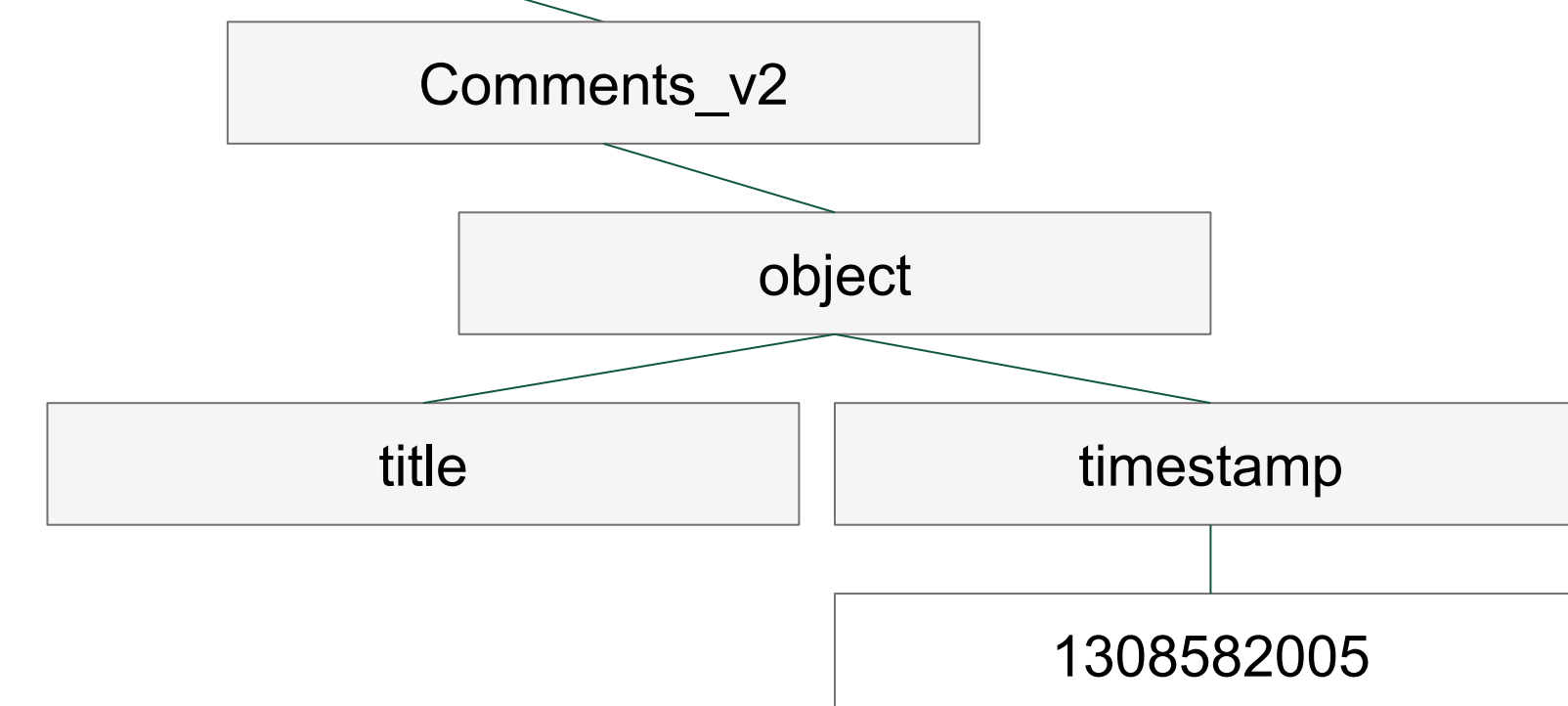
Iterative construction of a unified representation:



II: Tree Based on Semi-Structured Data



Implementation node	node type
folder	empty, folder
file	(semi-)structured, text, media, other
structure	object, array, key
leaf(element)	integer, datetime, mail, ..., string



Order-Aware Representation

preserving the structures:
concepts might have an order!

lists: ordered collections

maps: unordered collections

layer	order	cardinality	concept
folders	ordered	1:n	folder
file	ordered	1:n	structured, semi-structured, unstructured
structure	order-aware	n:m	list maps

Order-Aware Representation

most related work assumes either ordered or unordered structures

JSON tree¹:

order-aware JSON representation

no generalized representation yet

¹Hütter et al. (2022). JEDI: These aren't the JSON documents you're looking for... .
In: Proceedings of the 2022 International Conference on Management of Data,
ser. SIGMOD '22, Philadelphia, PA, USA: ACM, pp. 1584–1597.
DOI: 10.1145/3514221.3517850.

JEDI: These aren't the JSON documents you're looking for... (Extended Version*)

Thomas Hütter
Nikolaus Augsten
thomas.huetter@plus.ac.at
nikolaus.augsten@plus.ac.at
University of Salzburg
Austria

Christoph M. Kirsch
ck@cs.uni-salzburg.at
University of Salzburg
Austria
Czech Technical University
Czech Republic

Michael J. Carey
Chen Li
mjcarey@ics.uci.edu
chenli@ics.uci.edu
University of California, Irvine
USA

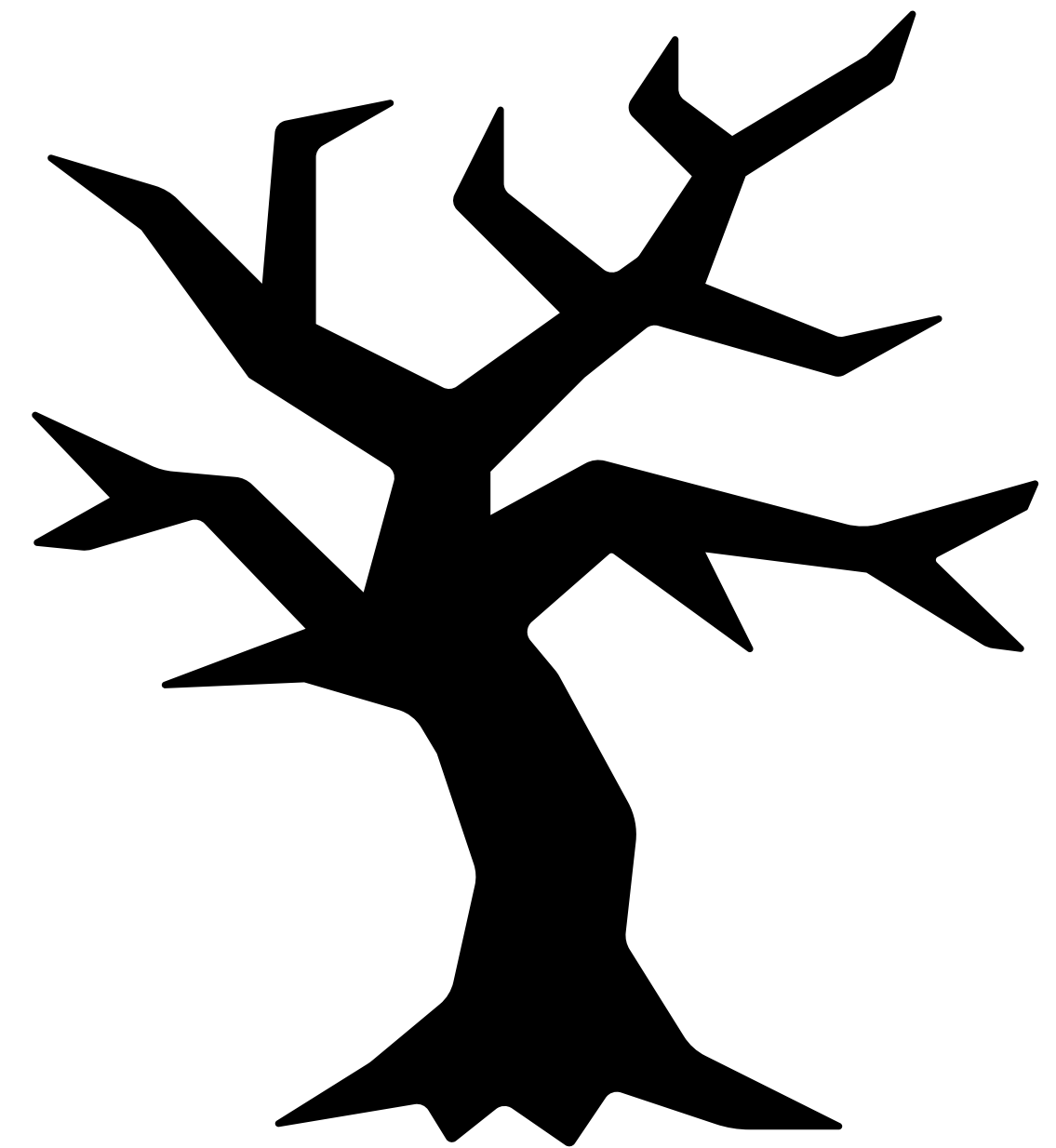
ABSTRACT

The JavaScript Object Notation (JSON) is a popular data format used in document stores to natively support semi-structured data. In this paper, we address the problem of JSON similarity lookup queries: given a query document and a distance threshold τ , retrieve all JSON documents that are within τ from the query document. Due

duplicates is not effective since key names or the structure will typically vary. Consider the document in Figure 1a, which was discovered by the crawler. The document in Figure 1b already resides in the database. Both documents represent the same movie, but due to their different representations a query for exact duplicates will fail. Detecting near-duplicate entries remains a challenge.

Tree-based SARPs

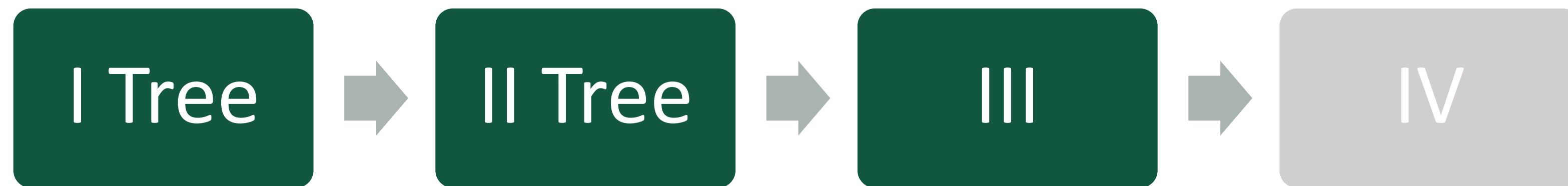
- + order-aware tree preserves the structures
- + easily query- and filterable
- this is nothing new information



Let's add some decorations!

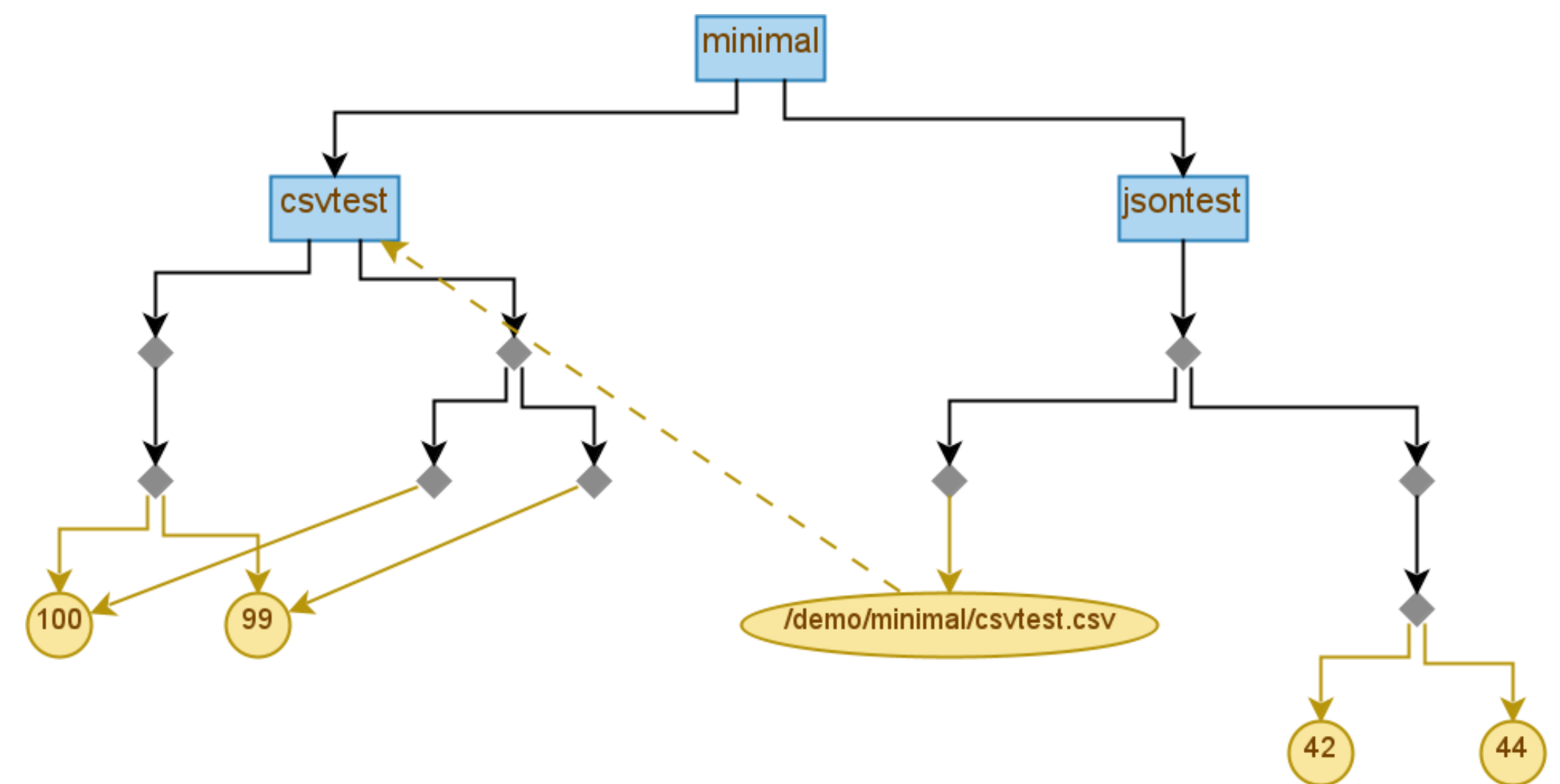
Approach

Iterative construction of a unified representation:



SARPs Convey Relations!

resolve relations
„factorization“

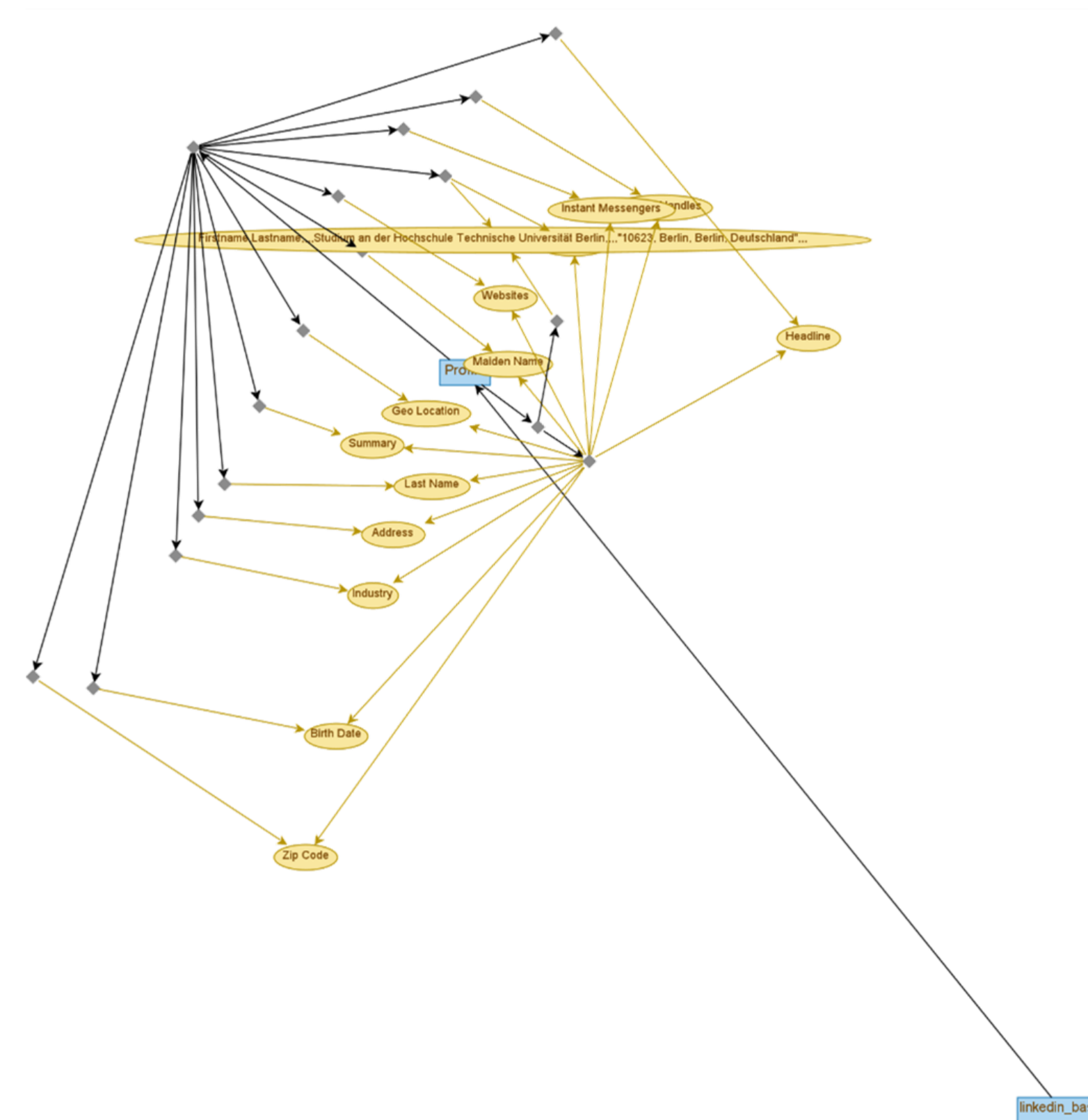


SARP Graph

„factorization“:

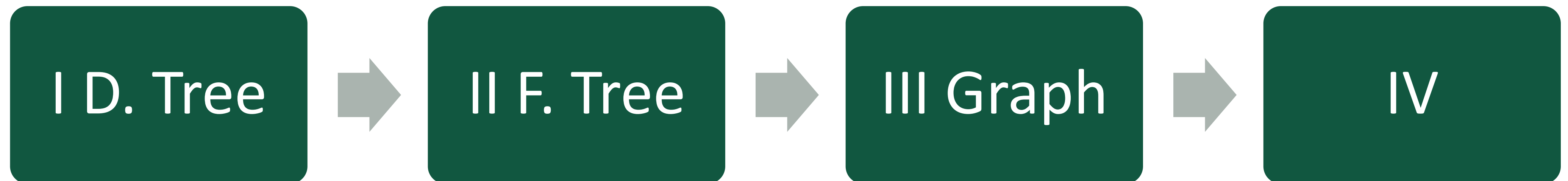
merge nodes that point to same value

→ more than 1 ingoing edge possible



Approach

Iterative construction of a unified representation:



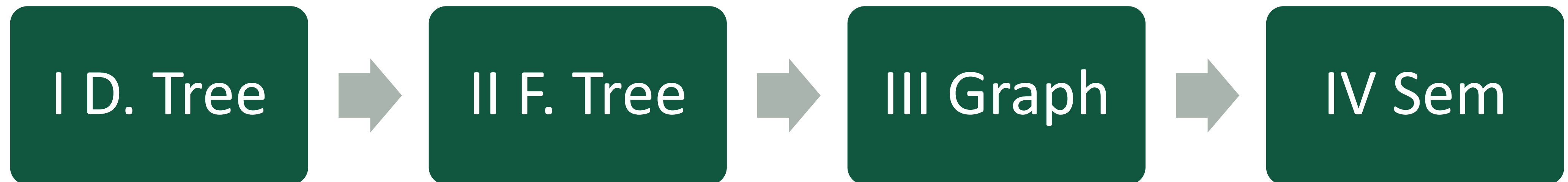
Semantic Enrichment

adding meta-data to make implicit knowledge explicit:

- fine-granular data types
 - pattern based detection: date, e-mail, etc.
 - NER
- describing structural nodes: statistical values
 - average length
 - number of child nodes
 - ...

Approach

Iterative construction of a unified representation:



Prototype

SARP graph
converter
→ I-II

SARP graph
analyzer
→ IV

proof-of-
concept
visualization

github



implemented in Java 17
NER with OpenNLP

Evaluation

Validation

minimal example

- every file format
- all constructs

Real-World Applicability

public data set:

- 2 data subjects
- 5 controllers each

Leschke, N., Pöhn, D., & Pallas, F. (2024). How to Drill into Silos: Creating a Free-to-Use Dataset of Data Subject Access Packages. In: Privacy Technologies and Policy (APF '24). pp. 132-155. Karlstad, Sweden. Springer, Cham.
https://doi.org/10.1007/978-3-031-68024-3_7



How to Drill into Silos: Creating a Free-to-Use Dataset of Data Subject Access Packages

Nicola Leschke¹✉, Daniela Pöhn², and Frank Pallas¹

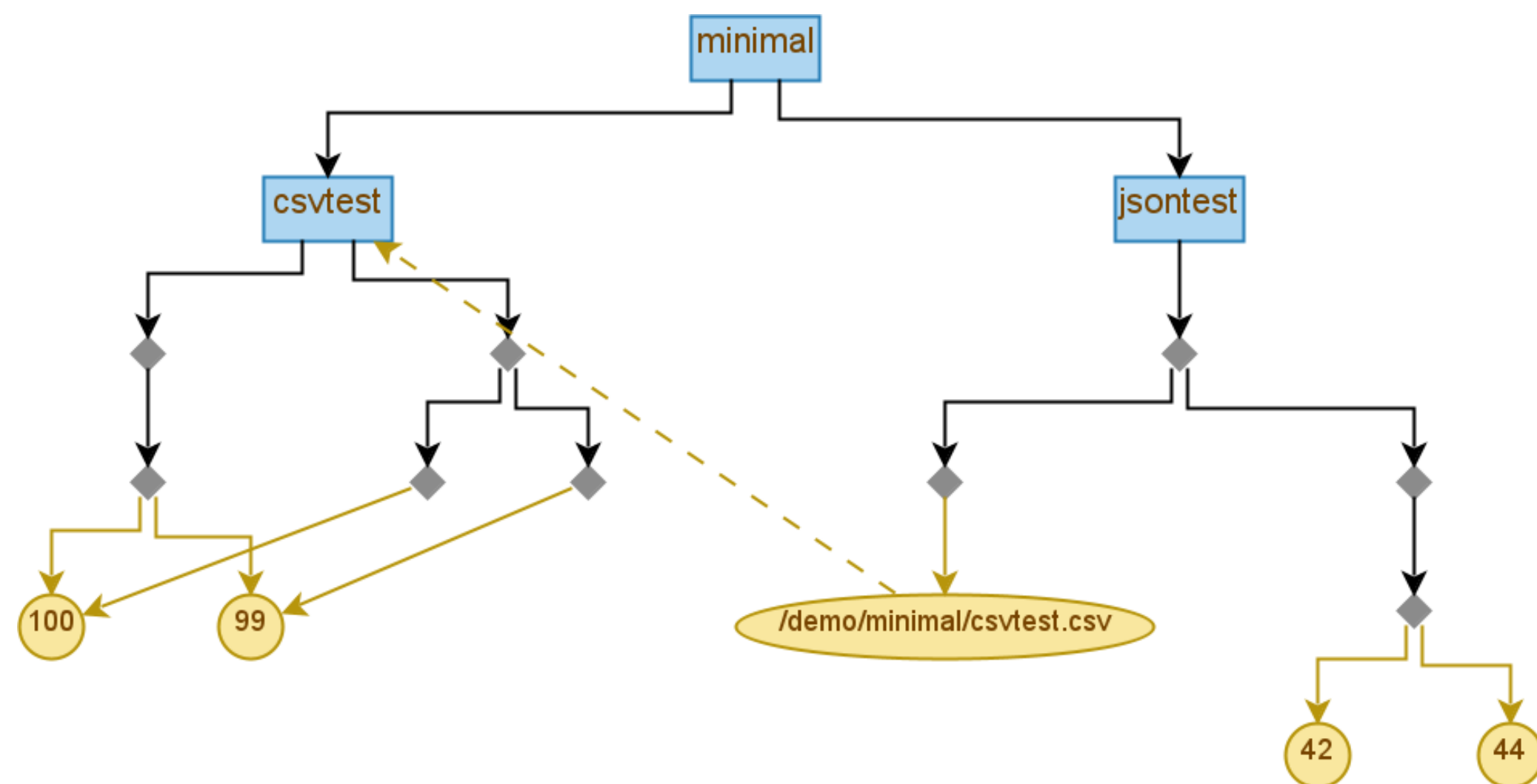
¹ Paris Lodron Universität Salzburg, Salzburg, Austria
{nicola.leschke, frank.pallas}@plus.ac.at

² Universität der Bundeswehr München, München, Germany
daniela.poehn@unibw.de

Abstract. The European Union's General Data Protection Regulation (GDPR) strengthened several rights for individuals (data subjects). One of these is the data subjects' right to access their personal data being collected by services (data controllers), complemented with a new right to data portability. Based on these, data controllers are obliged to provide respective data and allow data subjects to use them at their own discretion.

However, the subjects' possibilities for actually using and harnessing said data are severely limited so far. Among other reasons, this can be attributed to a lack of research dedicated to the actual use of controller-provided subject access request packages (SARPs). To open up and facilitate such research, we outline a general, high-level method for generating a public dataset of SARPs.

Evaluating SARP Graph



no connections – why?

Finding: SARPs are broken!

```
{
  "sender_name": "Subject B",
  "timestamp_ms": 1709660406201,
  "gifs": [
    {
      "uri":
"your_activity_across_facebook/messages/inbox/subject_b/gifs/
364927764_1177536576972456_8237558561056084391_n_799975290337
4705.gif"
      your_activity_across_facebook/messages/inbox/subject_b_122128201010107929/gifs
/364927764_1177536576972456_8237558561056084391_n_7999752903374705.gif
    },
    "is_geoblocked_for_viewer": false
  ],
}
```

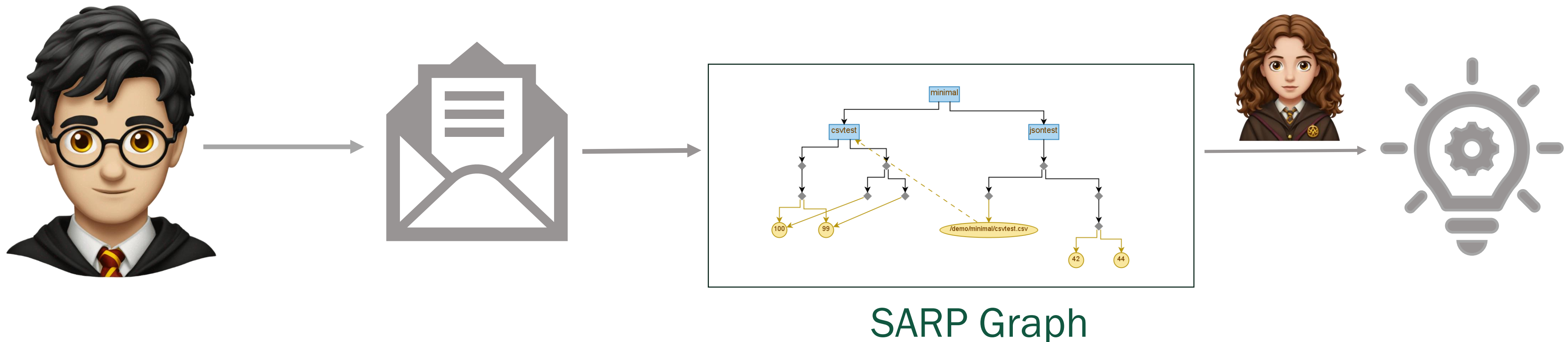


Conclusion

SARP-Graph

- We proposed an iterative approach for a unified, machine-readable representation of SARPs by means of graph inference.
- We provide a prototypical converter open source.
- Evaluation shows that real-world SARPs are mal-formed.

Coming back to Harry....

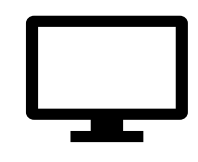


SARP Graph shall not be the end-product, but an enabler for integrated SARP-based solutions → future work.

Alohomora! Facilitating Personal Data Access Through Automated Graph Extraction and Integration

Nicola Leschke

 nicola.leschke@plus.ac.at

 <https://pepsys-group.github.io/>

 Nicola Leschke

This work has been supported by the EXDIGIT (Excellence in Digital Sciences and Interdisciplinary Technologies) project, funded by Land Salzburg under grant number 20204-WISS/263/6-6022.