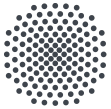
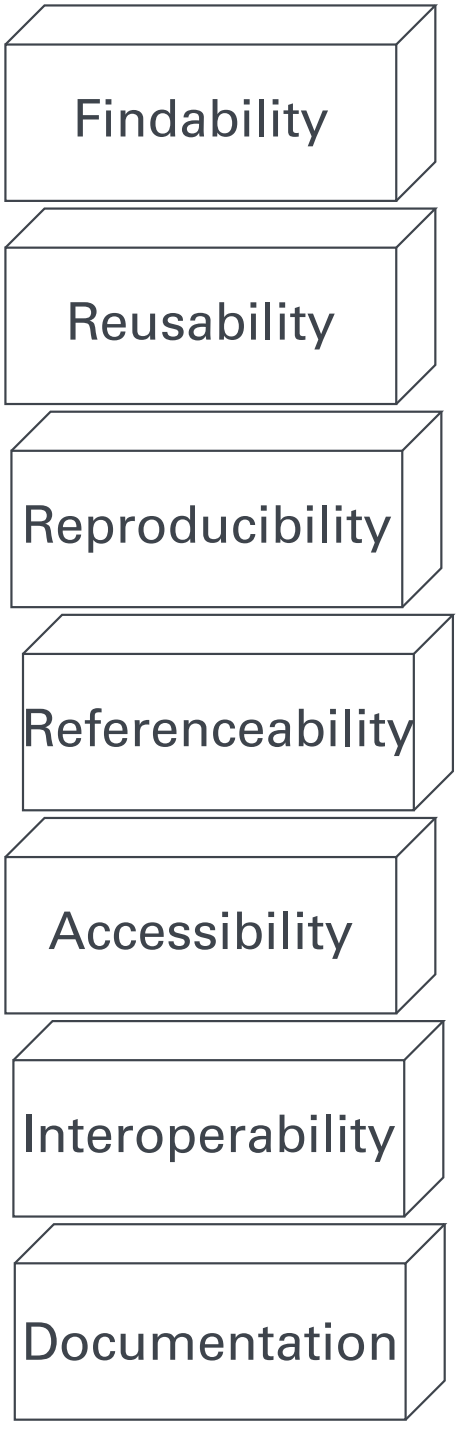


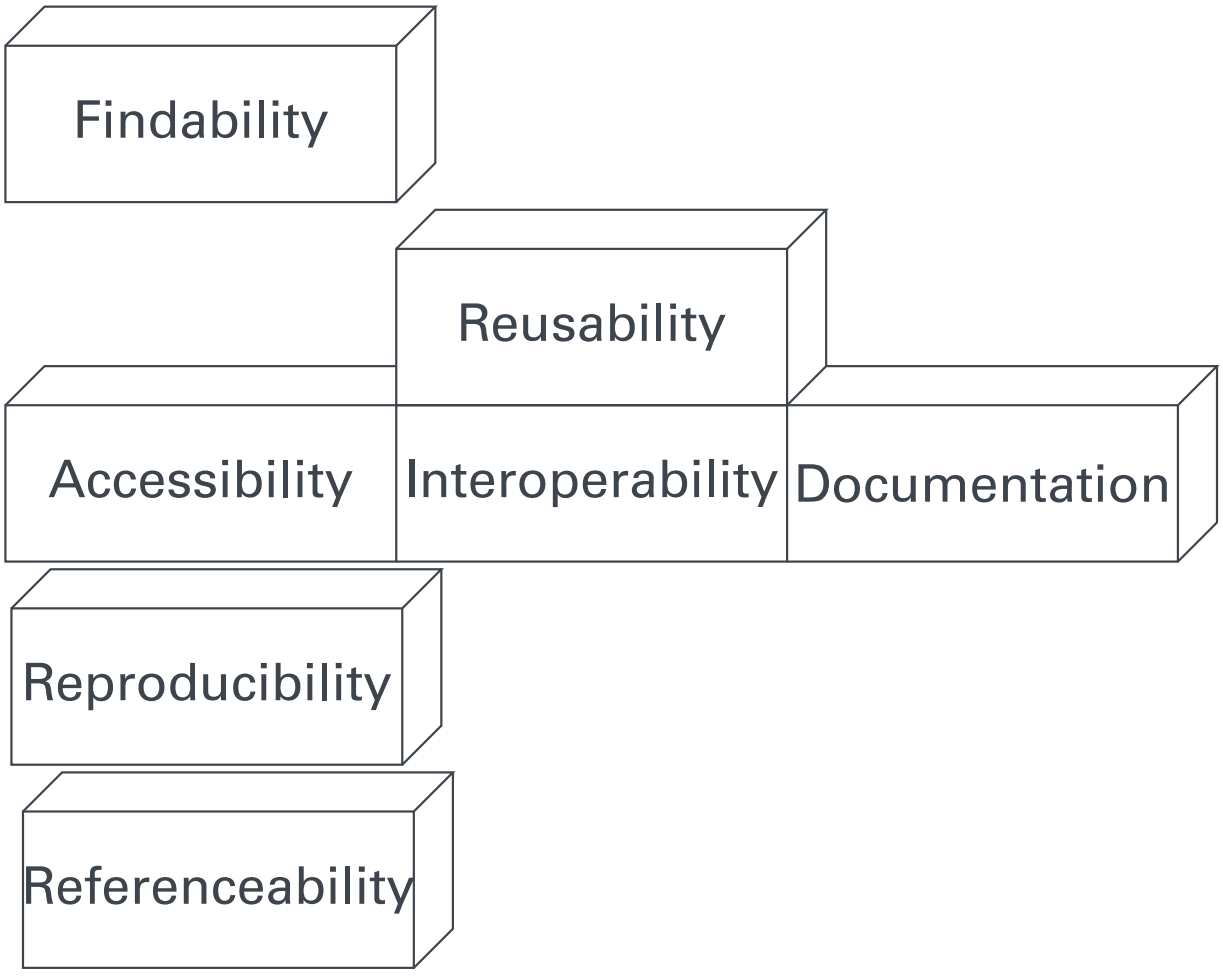
Universität Stuttgart
Institut für Maschinelle
Sprachverarbeitung

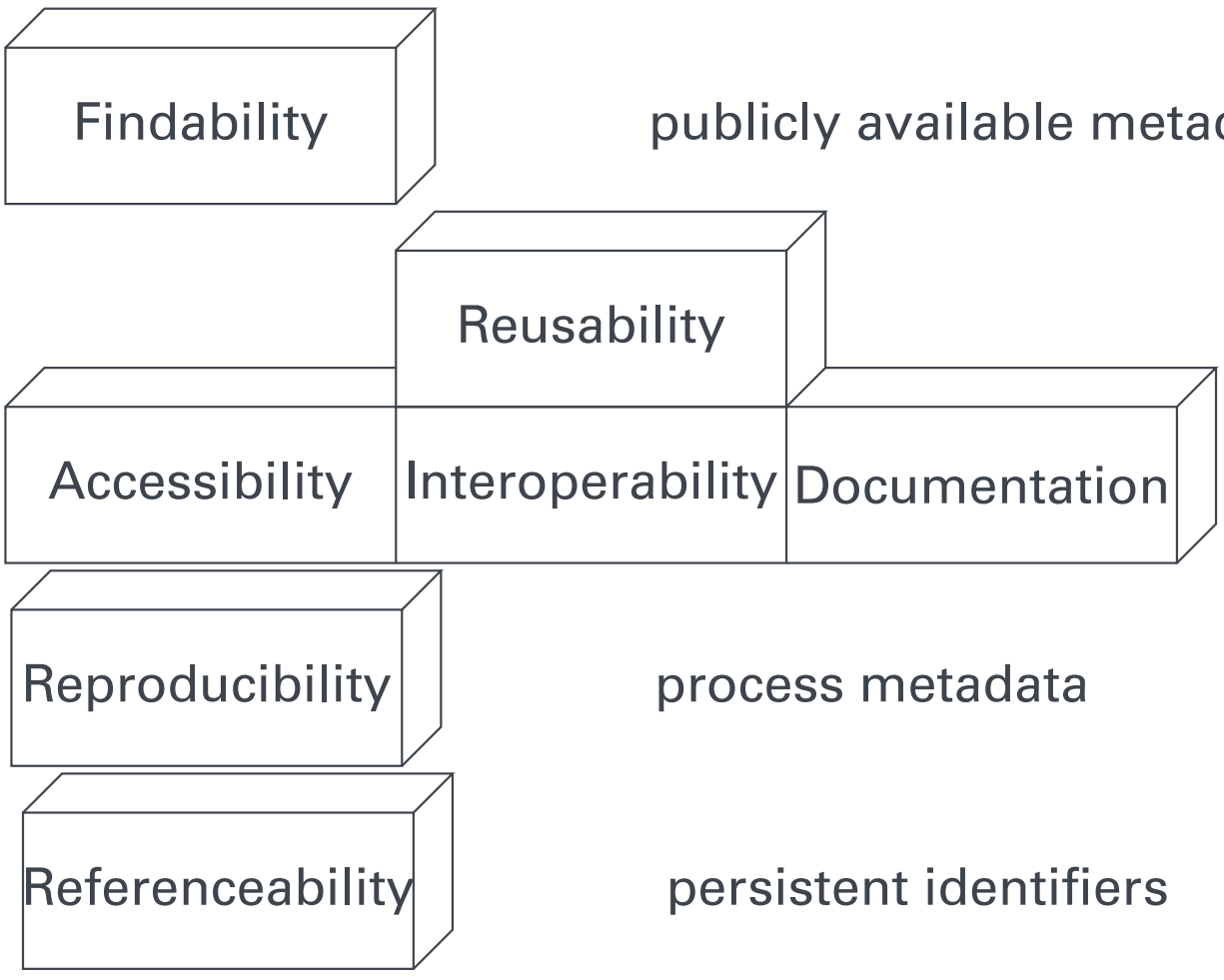
Kerstin Jung,
Markus Gärtner

Approaches to Sustainable Process Metadata

Aspects of Sustainability


Findability
Reusability
Reproducibility
Referenceability
Accessibility
Interoperability
Documentation


Findability
Reusability
Accessibility
Interoperability
Documentation
Reproducibility
Referenceability


Findability
publicly available metadata
Reusability
Accessibility
Interoperability
Documentation
Reproducibility
process metadata
Referenceability
persistent identifiers

Metadata vs. Process Metadata

Metadata …

- describes a resource
- is not part of the resource
- is part of the documentation of a resource
- comprises search terms for potential users
- provides starting point to reuse:
  “Does this resource fulfill my requirements?”

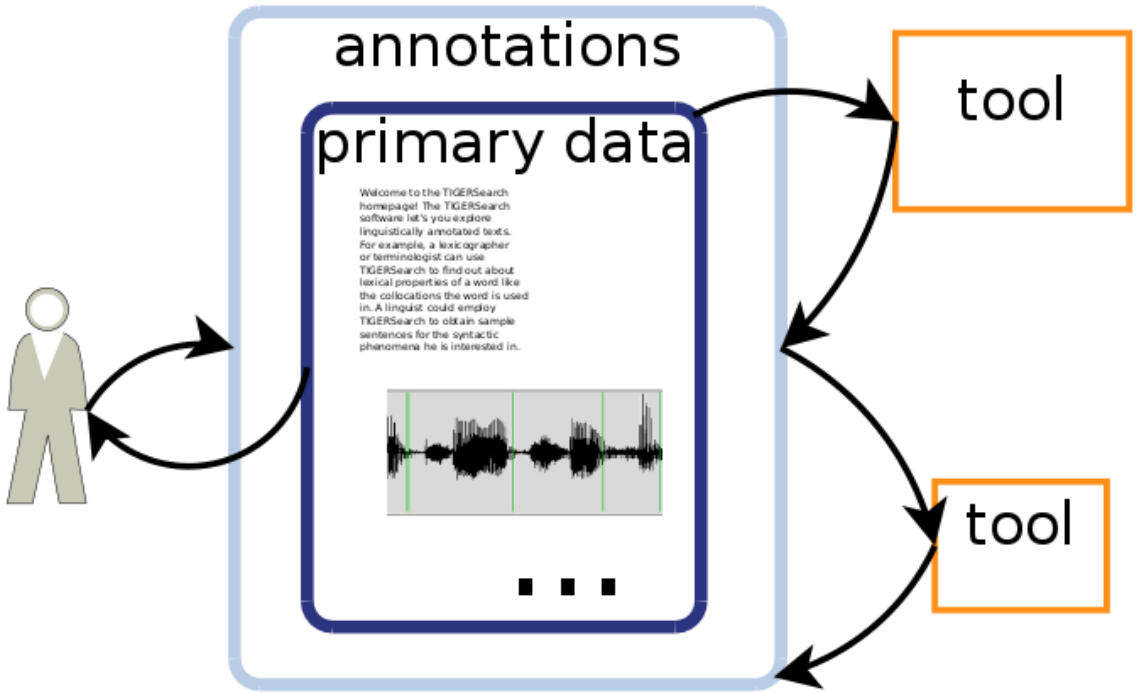
Metadata should …

- exist in a human readable form
- be automatically searchable
- be able to describe different types of resources
- be aware of different user groups

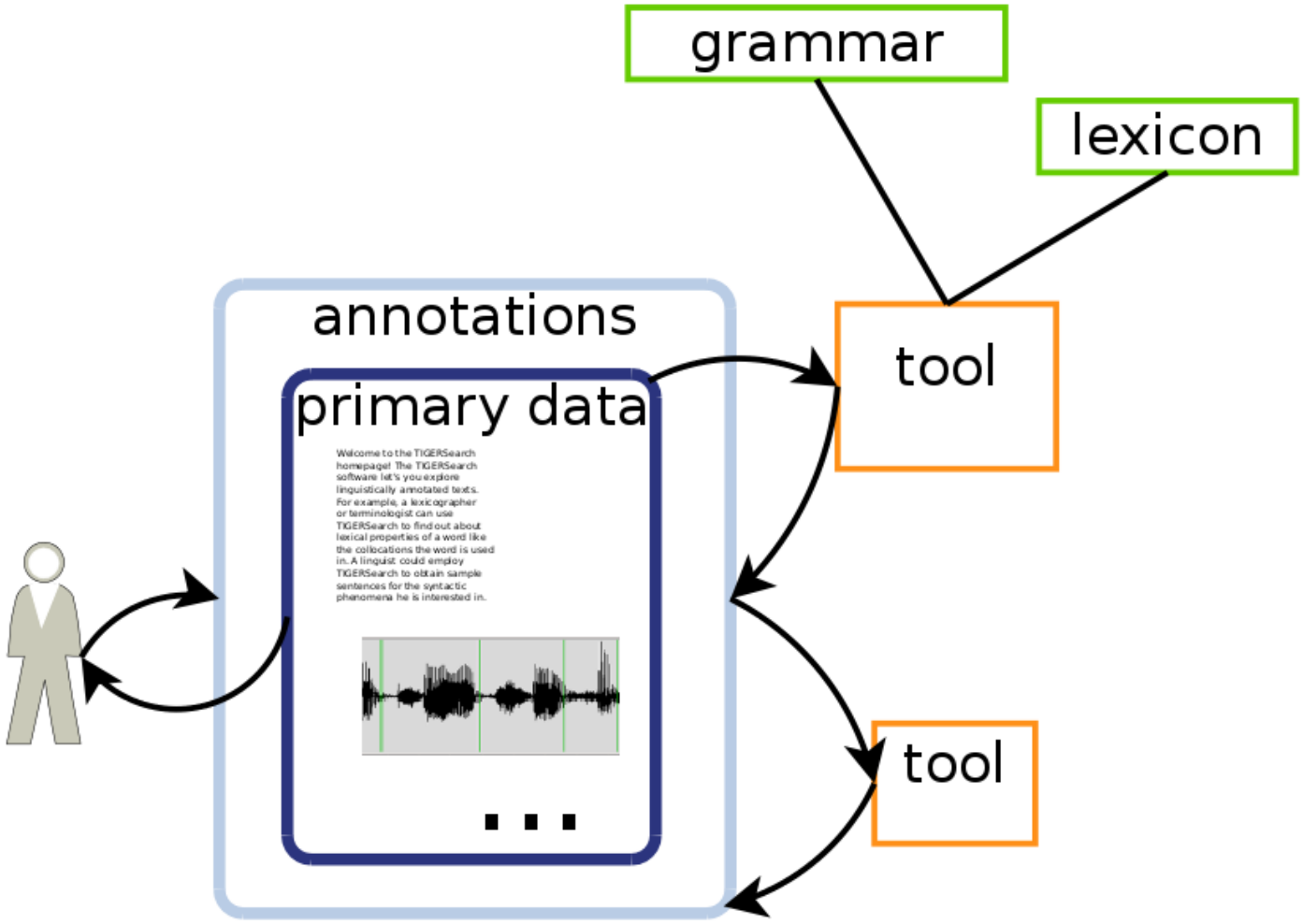
Detour: Resources

Detour: Resources

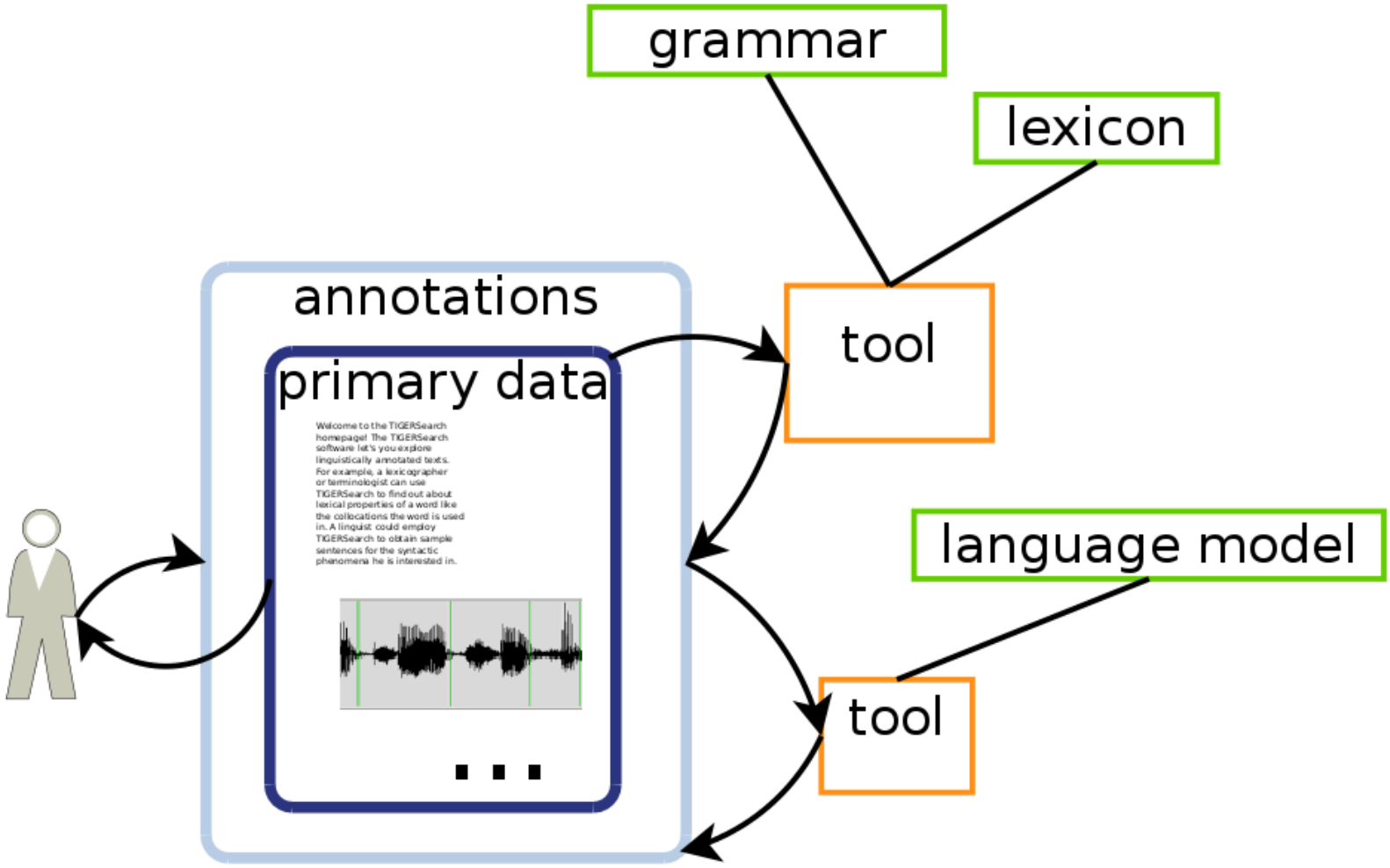
Detour: Resources



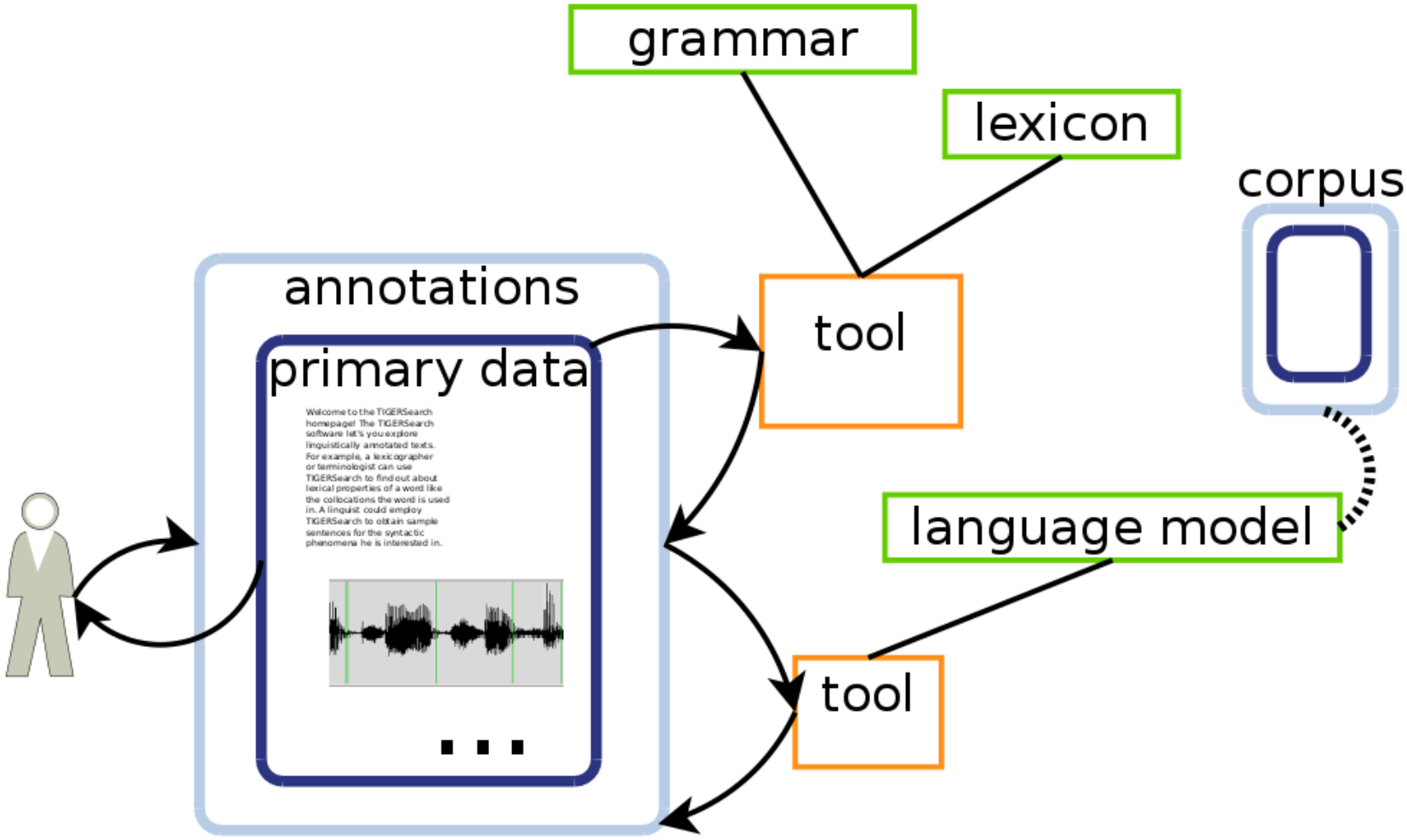
Detour: Resources



Detour: Resources



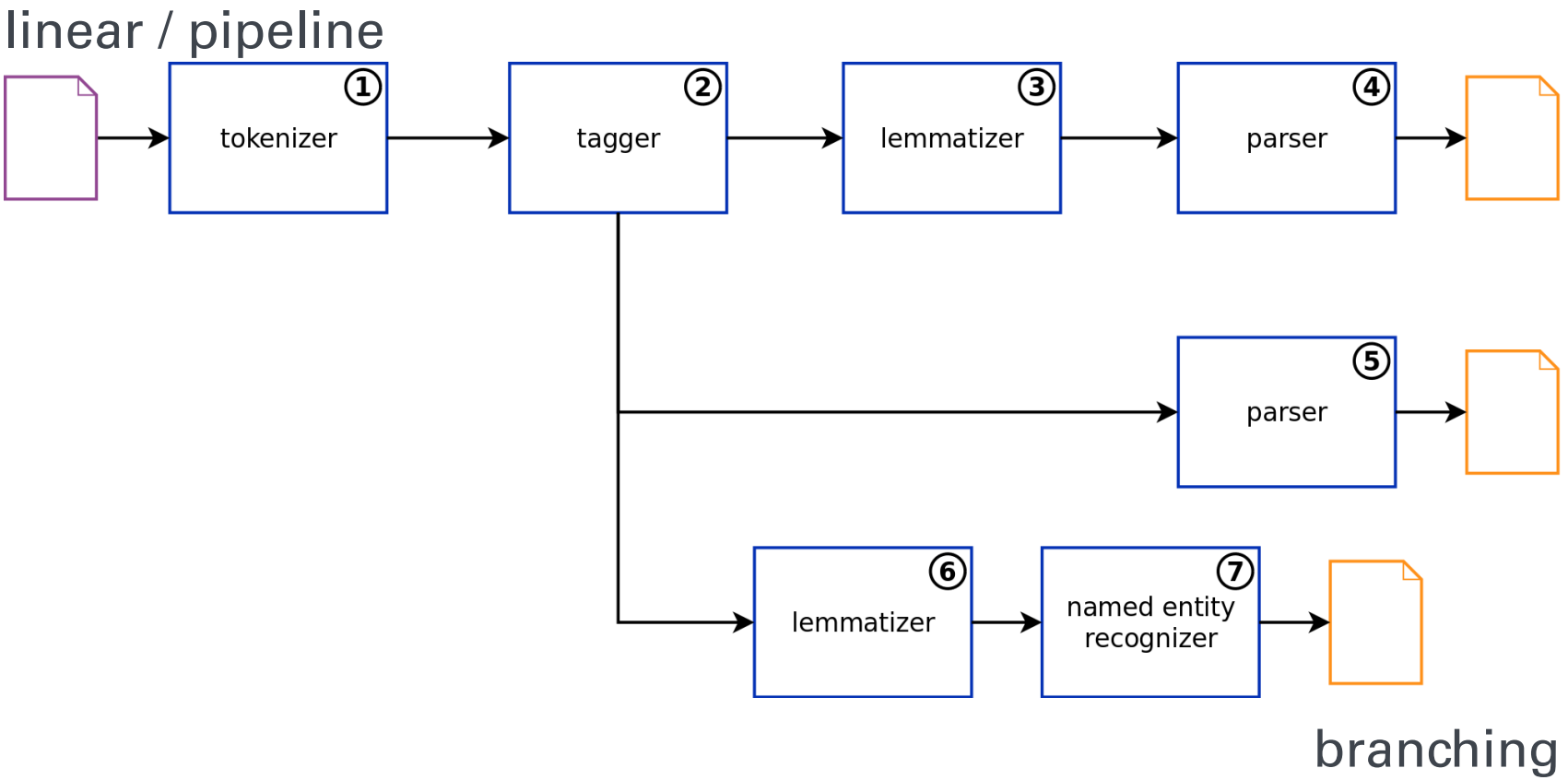
Detour: Resources



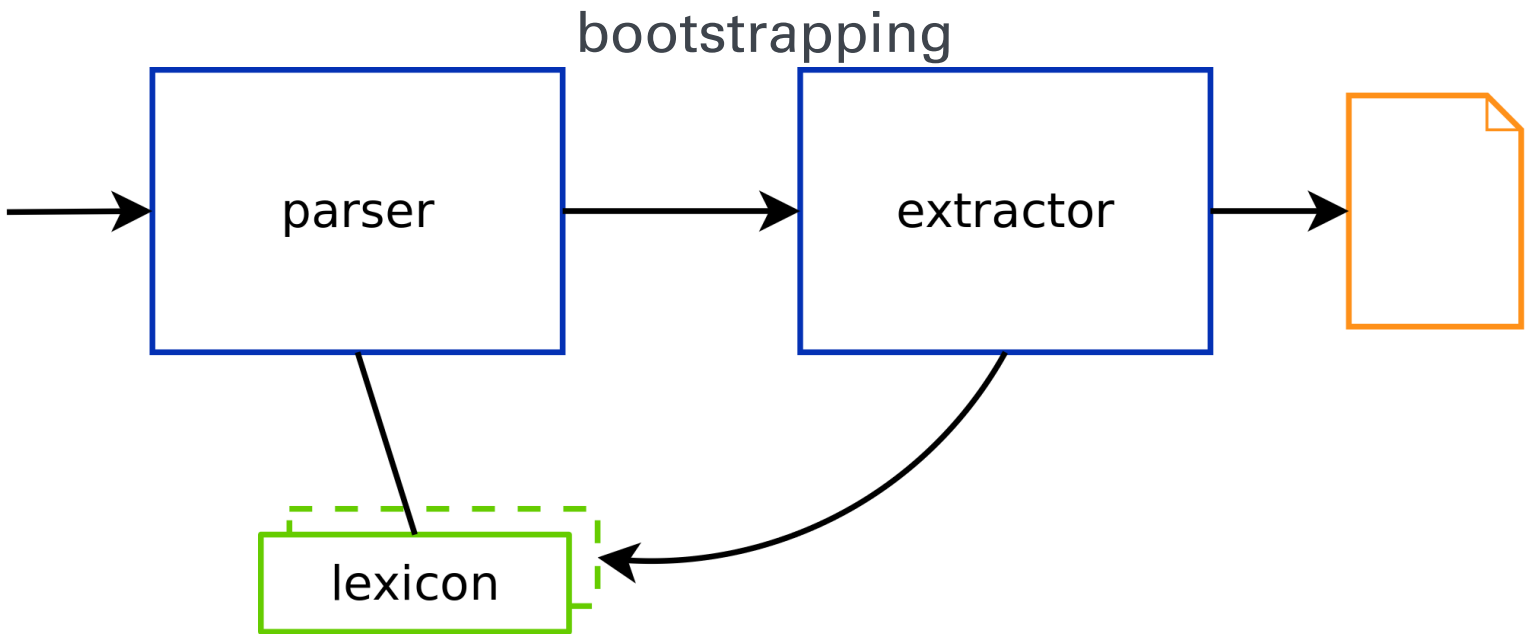
Process metadata …

- is data on the creation of a resource or the process of a study
  $\Rightarrow$ documents a workflow
- contains important information for the documentation of a resource
- provides information regarding reuse:
  “Does this resource fulfill my requirements?”
- enhances reproducibility and comparability:
  “Which tool version was applied on which subcorpus?”
  “Which type of speakers was used for training the subjects in the study?”
- needs to represent manual as well as automatic processing steps

Detour: Workflow types and analysis relations



Detour: Workflow types and analysis relations



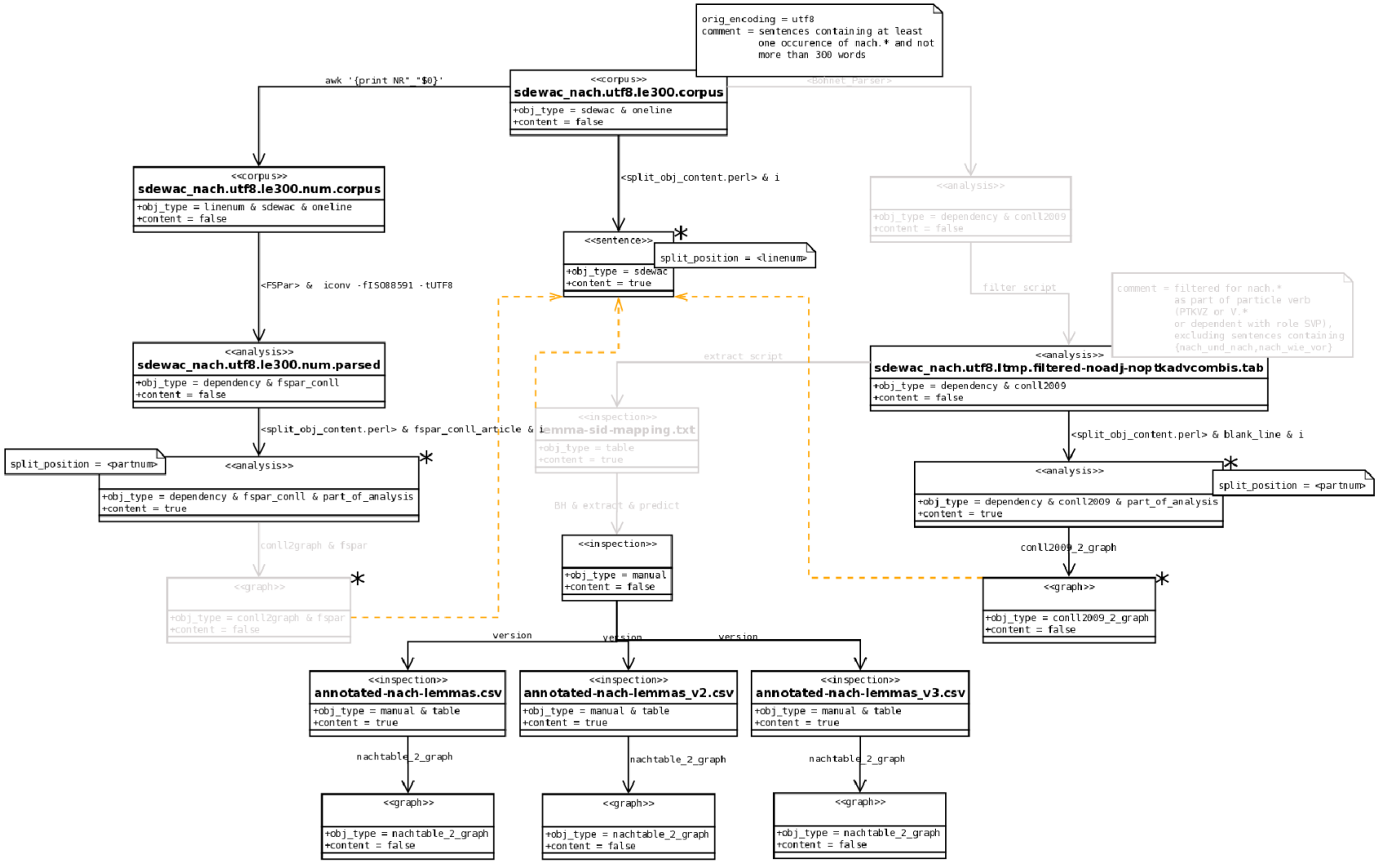
Detour: Workflow types and analysis relations

- Vertical relations
  - pipeline threads
  - analyses are based on output of preceding workflow steps
- Horizontal relations
  - concurrent analyses of the same layer of description
  - independent of each other
  - comparability based on starting point
- Temporal relations
  - development of resources
    e.g. different versions of the same tool,
    applied on the same set of input data
  - hypothesis on quality development

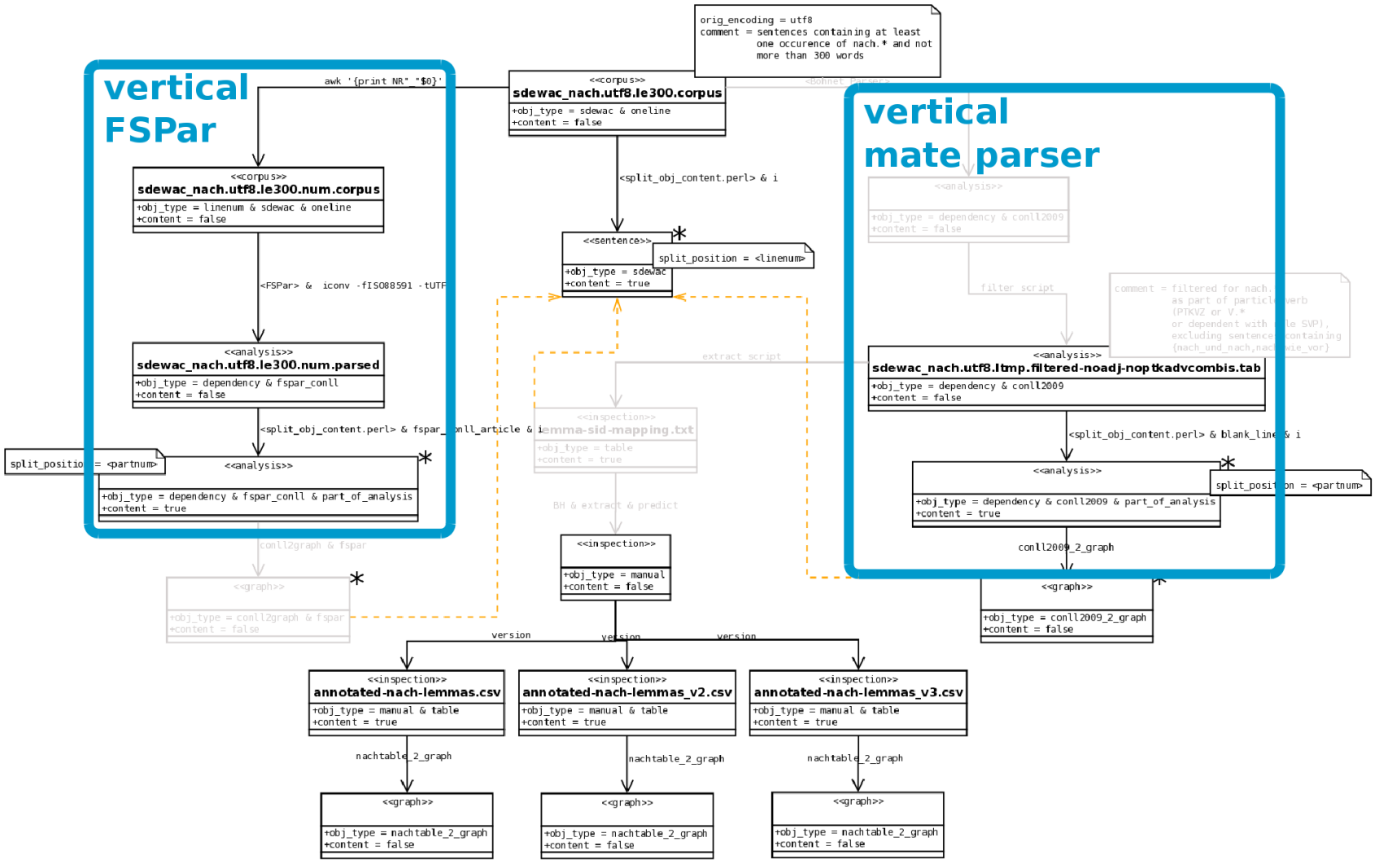
Detour: Workflow types and analysis relations

- Vertical relations
  - pipeline threads
  - analyses are based on output of preceding workflow steps
- Horizontal relations
  - concurrent analyses of the same layer of description
  - independent of each other
  - comparability based on starting point
- Temporal relations
  - development of resources
    e.g. different versions of the same tool,
    applied on the same set of input data
  - hypothesis on quality development

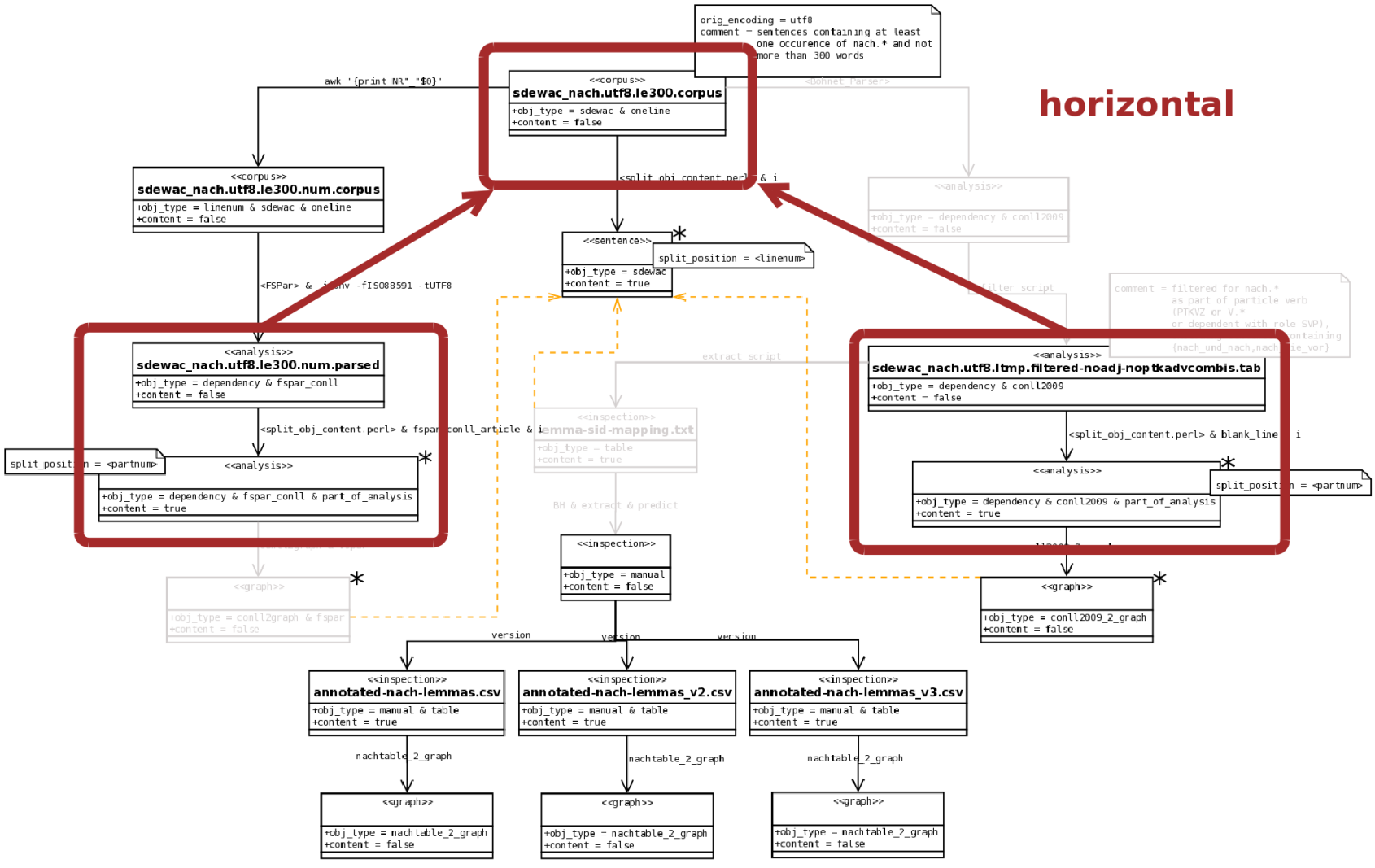
Detour: Workflow types and analysis relations



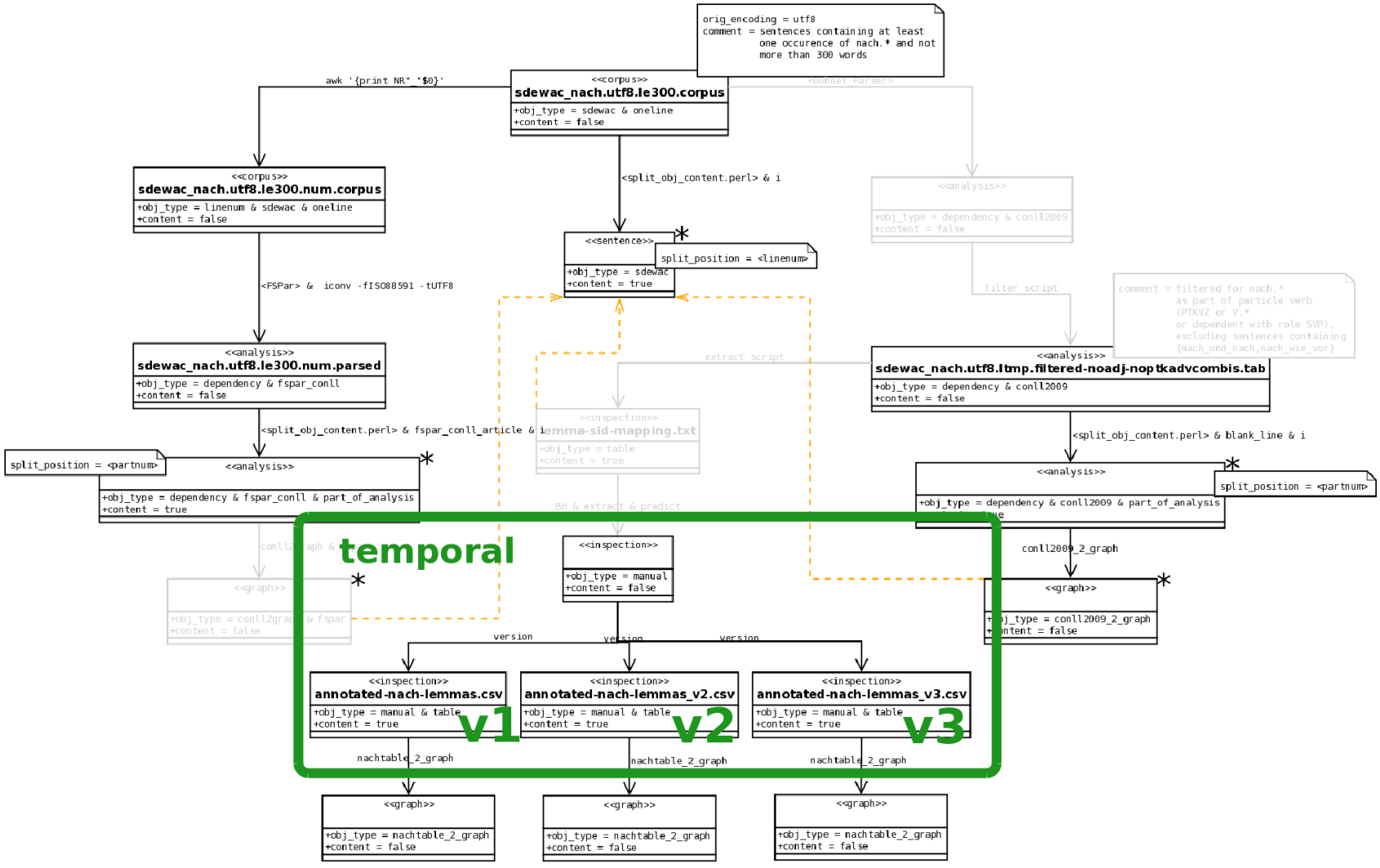
Detour: Workflow types and analysis relations



Detour: Workflow types and analysis relations



Detour: Workflow types and analysis relations



Metadata: Examples

Dublin Core Metadata Element Set, Version 1.1

- contributor
- coverage
- creator
- date
- description
- format
- identifier
- language
- publisher
- relation
- rights
- source
- subject
- title
- type

Metadata: Examples

- TEI Header
- Text Encoding Initiative
  P5: Guidelines for Electronic Text Encoding and Interchange

  fileDesc bibliographic description: title, author, editor, …
  encodingDesc tools, normalizations, …
  profileDesc text profile: language, text topic, …
  revisionDesc steps of revision: annotations, corrections, …

`http://www.tei-c.org/release/doc/tei-p5-doc/de/html/HD.html`

Metadata: Examples

- TEI Header – LAUDATIO Repository
- Long-term Access and Usage of Deeply Annotated Information

Lühr, Rosemarie; Faßhauer, Vera; Prutscher, Daniela; Seidel, Henry; Fuerstinnenkorrespondenz (Version 1.1), Universität Jena, DFG.

```
<teiHeader type="CorpusHeader">
−<fileDesc>
  +<titleStmt></titleStmt>
   <extent type="Tokens">262.465</extent>
  −<publicationStmt>
     <authority>Universität Jena, DFG</authority>
    −<idno>
       DFG-Projekt "Frühneuzeitliche Fürstinnenkorrespondenz im mitteldeutschen Raum"
     </idno>
    +<availability status="free"></availability>
     <date when="2013-11" type="CorpusRelease">Erste Veröffentlichung des Korpus.
     </date>
    −<date when="2014-09-05" type="CorpusRelease">
       Zweite Veröffentlichung des Korpus mit der Version 1.1.
     </date>
   </publicationStmt>
  +<sourceDesc></sourceDesc>
 </fileDesc>
```

http://hdl.handle.net/11022/0000-0000-2E44-1

Metadata: Examples

- TEI Header – LAUDATIO Repository
- Long-term Access and Usage of Deeply Annotated Information

Lühr, Rosemarie; Faßhauer, Vera; Prutscher, Daniela; Seidel, Henry; Fuerstinnenkorrespondenz (Version 1.1), Universität Jena, DFG.

```
+<profileDesc></profileDesc>
+<encodingDesc n="1"></encodingDesc>
+<encodingDesc n="2"></encodingDesc>
+<encodingDesc n="3"></encodingDesc>
−<revisionDesc>
    <change n="1.0" when="2013-11" who="NA" type="NA">NA</change>
  −<change n="1.1" when="2014-09-05" who="CorpusEditor" type="Konsistenzprüfung">
      Einzelne Korrekturen der Metadaten, der Annotationen und der Zuweisungen im
      gesamten Korpus. Keine Änderungen der Annotationsschemata und der
      Verarbeitsungsschritte.
    </change>
  </revisionDesc>
</teiHeader>
```

http://hdl.handle.net/11022/0000-0000-2E44-1

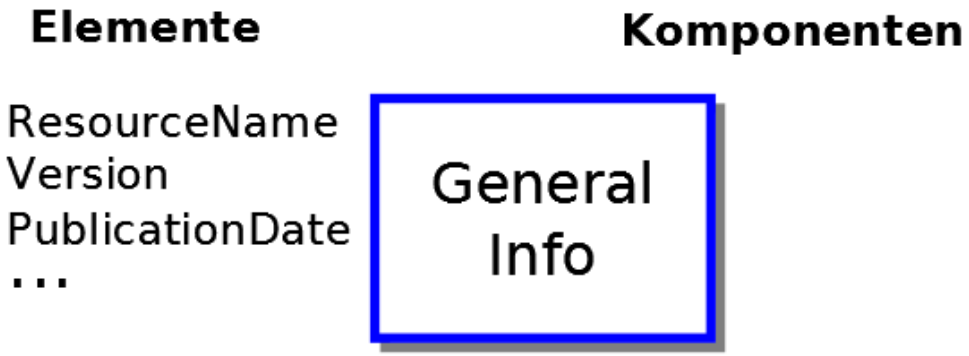
Metadata: Examples

- CLARIN: Component Metadata Infrastructure – CMDI
- components and profiles in Component Registry

Elemente **Komponenten**

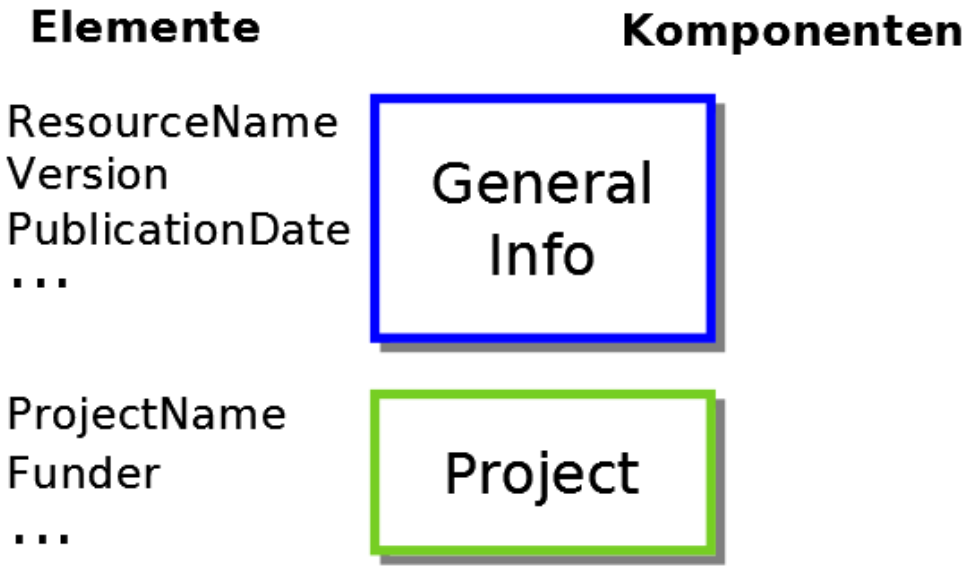
Metadata: Examples

- CLARIN: Component Metadata Infrastructure – CMDI
- components and profiles in Component Registry



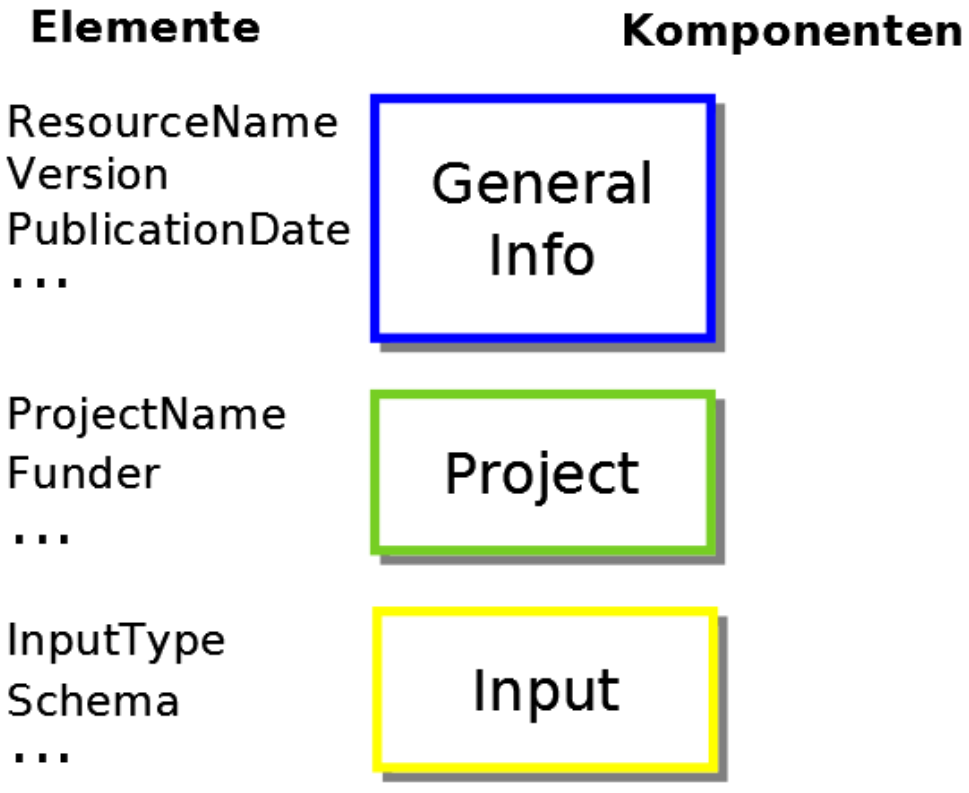
Metadata: Examples

- CLARIN: Component Metadata Infrastructure – CMDI
- components and profiles in Component Registry



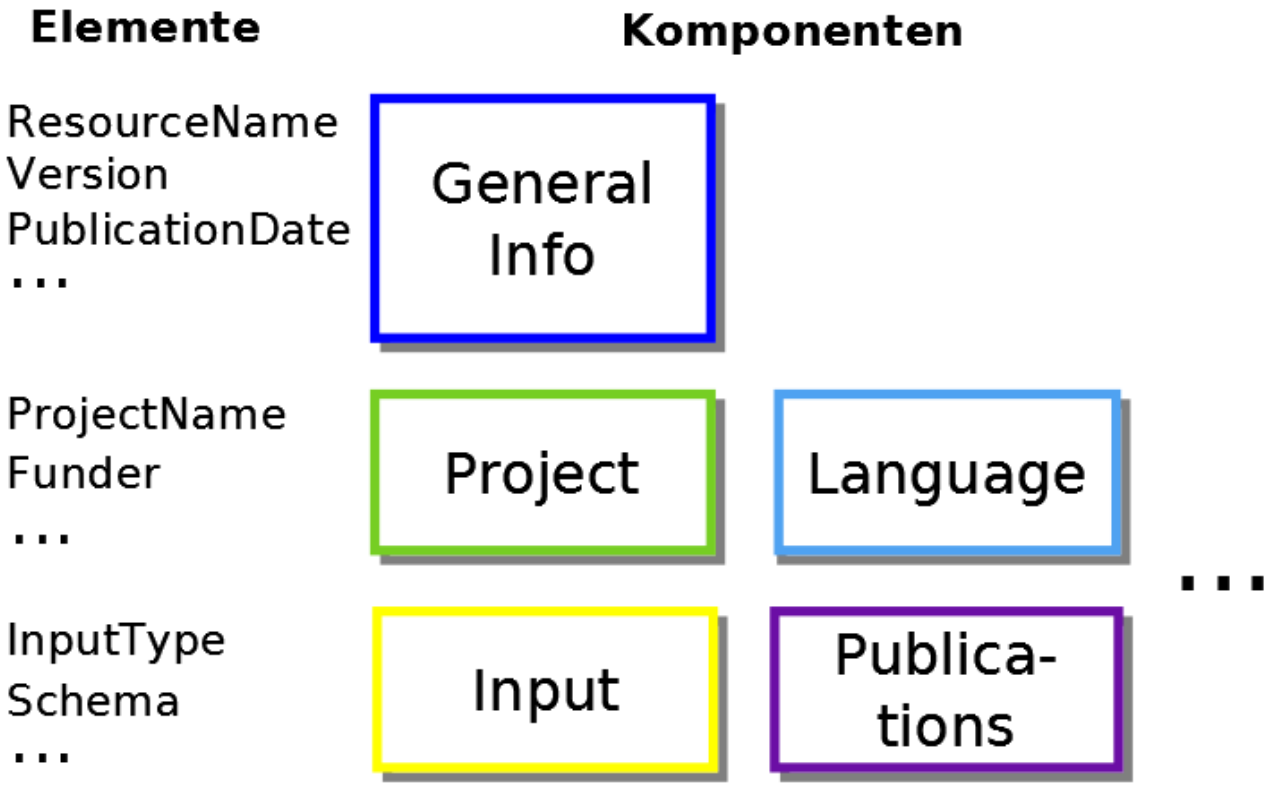
Metadata: Examples

- CLARIN: Component Metadata Infrastructure – CMDI
- components and profiles in Component Registry



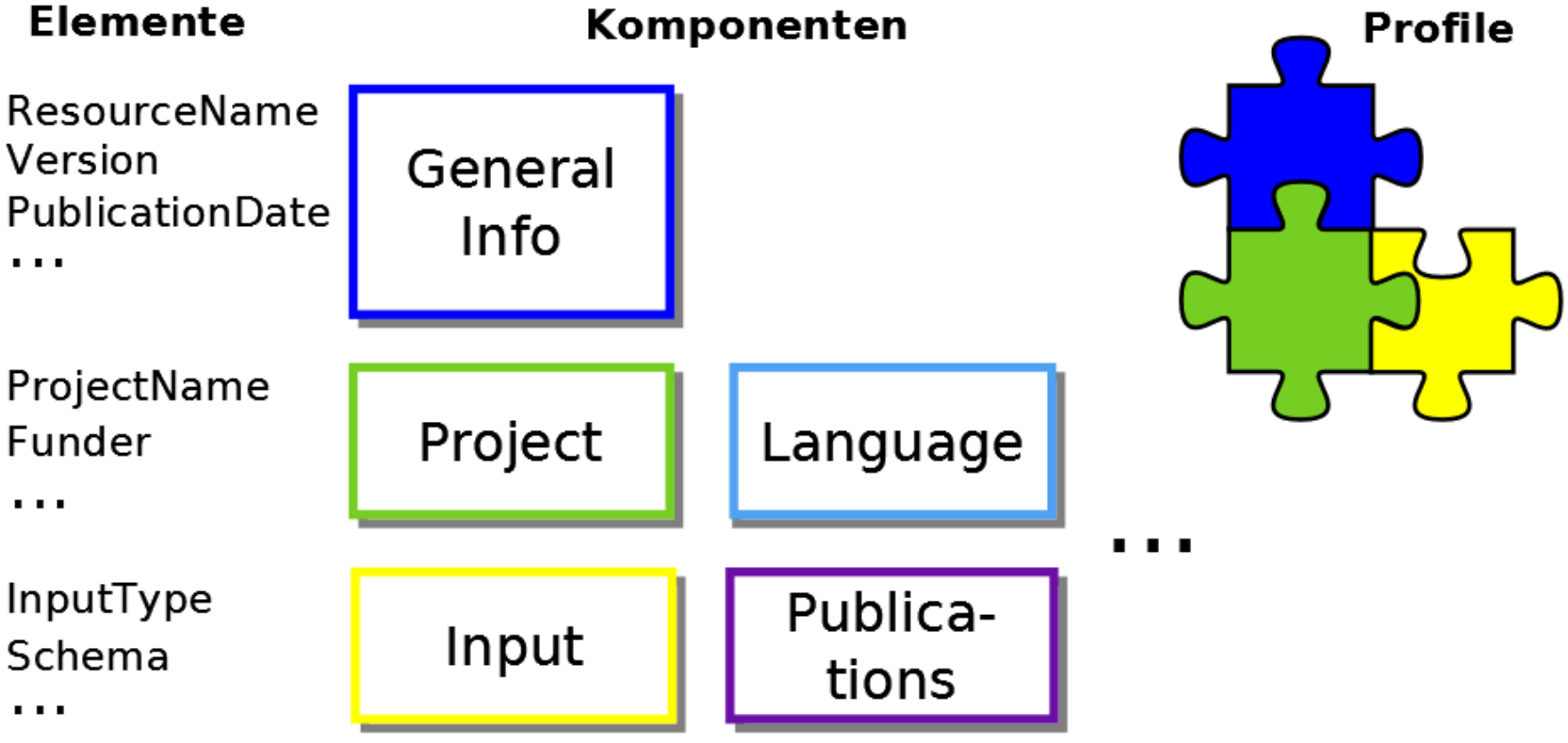
Metadata: Examples

- CLARIN: Component Metadata Infrastructure – CMDI
- components and profiles in Component Registry



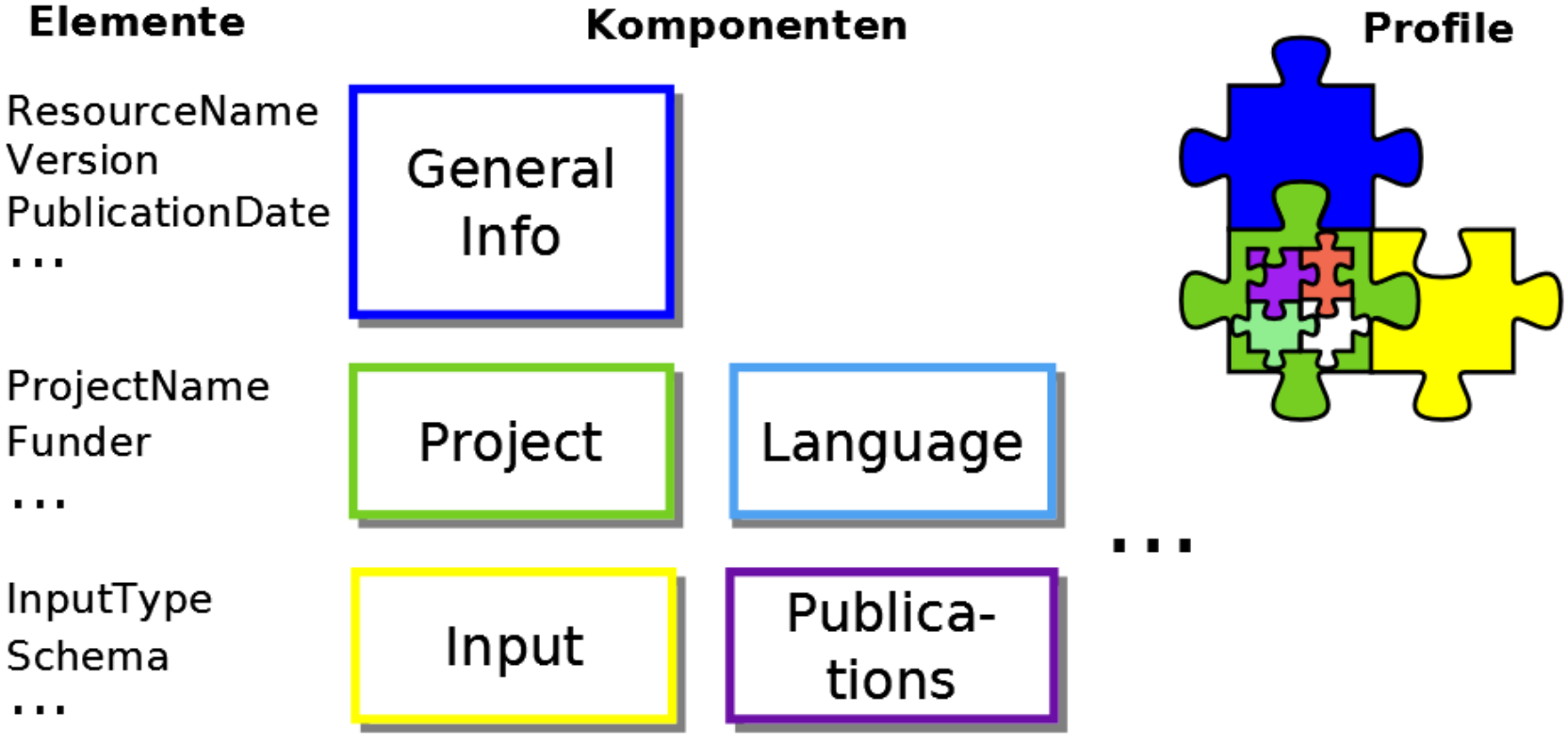
Metadata: Examples

- CLARIN: Component Metadata Infrastructure – CMDI
- components and profiles in Component Registry



Metadata: Examples

- CLARIN: Component Metadata Infrastructure – CMDI
- components and profiles in Component Registry



Metadata: Examples

- CLARIN: Component Metadata Infrastructure – CMDI
- components and profiles in Component Registry

```
<cmd:CreationToolInfo>
  <cmd:CreationTool>XLE system</cmd:CreationTool>
  <cmd:ToolType>annotation tool</cmd:ToolType>
</cmd:CreationToolInfo>
<cmd:CreationToolInfo>
  <cmd:CreationTool>TigerRegistry</cmd:CreationTool>
  <cmd:ToolType>converter</cmd:ToolType>
</cmd:CreationToolInfo>
<cmd:CreationToolInfo>
  <cmd:CreationTool>Salto</cmd:CreationTool>
  <cmd:ToolType>manual annotation support</cmd:ToolType>
  <cmd:ToolType>annotation manager</cmd:ToolType>
</cmd:CreationToolInfo>
```

http://hdl.handle.net/11858/00-247C-0000-0022-F7B1-A

Metadata: Examples

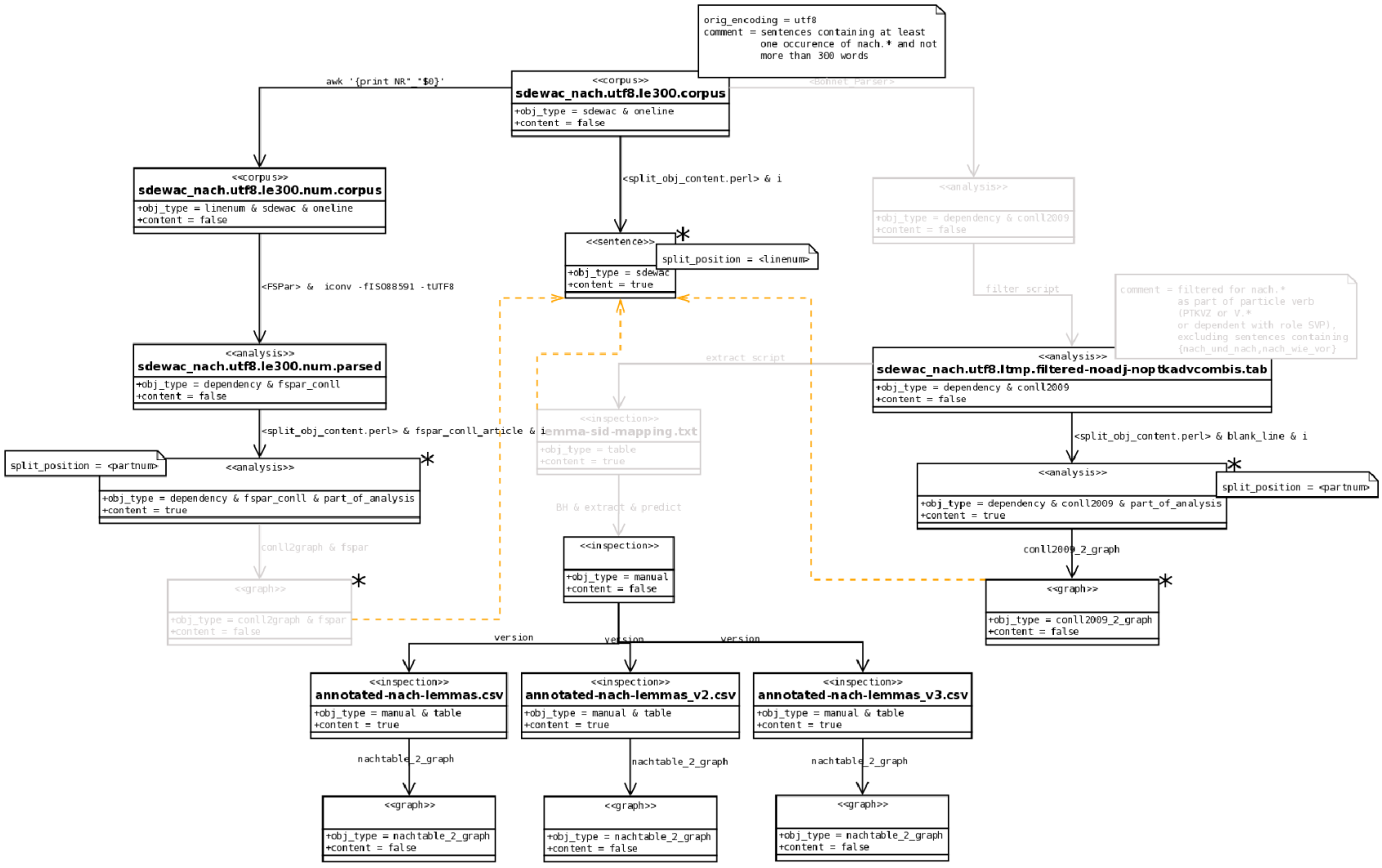
- CLARIN: WebLicht
- building web service chains – process metadata

```
<cmd:Toolchain>
−<cmd:ToolInChain>
  −<cmd:PID>
      http://hdl.handle.net/11858/00-1778-0000-0004-BA56-7
    </cmd:PID>
    <cmd:Parameter name="lang" value="de"/>
    <cmd:Parameter name="type" value="text/plain"/>
  </cmd:ToolInChain>
−<cmd:ToolInChain>
  −<cmd:PID>
      http://hdl.handle.net/11858/00-247C-0000-0007-3736-B
    </cmd:PID>
    <cmd:Parameter name="lang" value="de"/>
  </cmd:ToolInChain>
```

Keeping Track of the Workflow

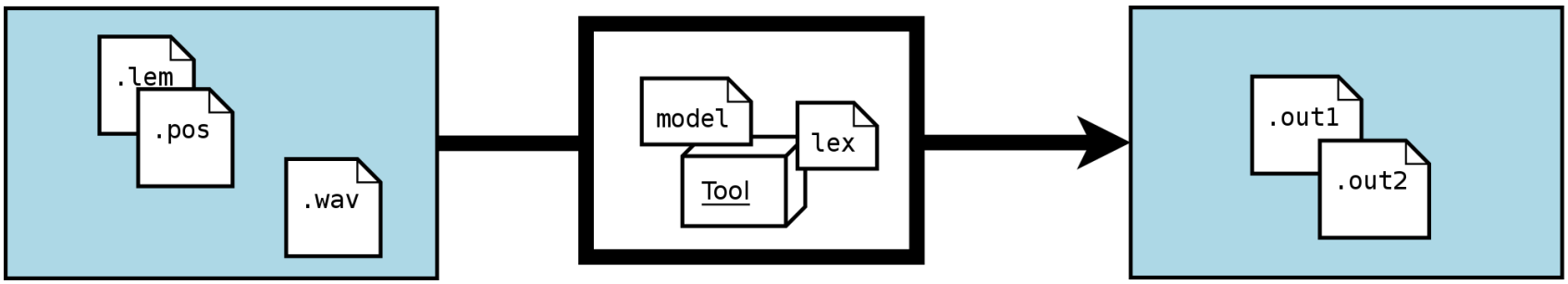
Workflow steps

Workflow steps



Workflow steps

Workflow steps



Tracking workflow steps as triples

- GRAIN corpus of German Radio Interviews
  - JSON-style process metadata schema
  - several annotation layers created by different projects
- RePlay-DH Client
  - JSON-style process metadata schema
  - supports elicitaion of process metadata

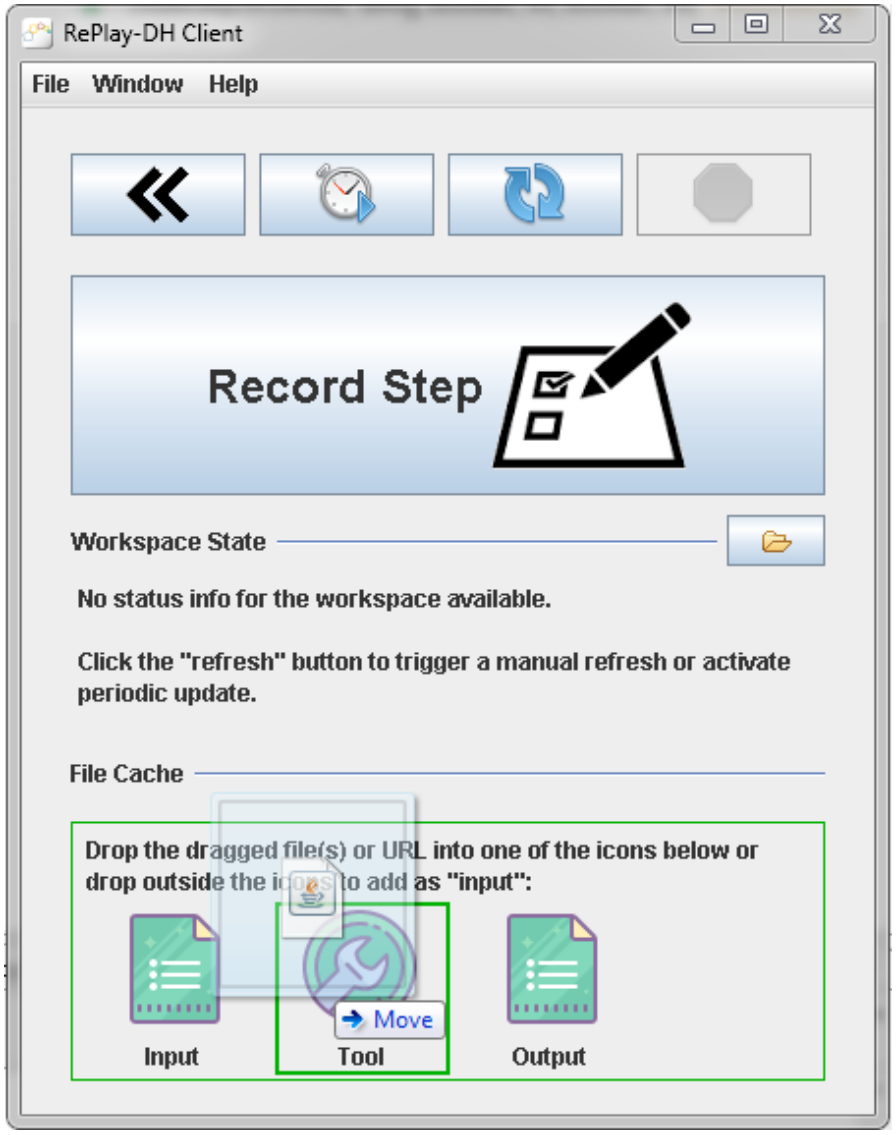
Elicitation of process metadata

- Documentation ahead of time: workplan, project description, …
  - details might not be known
  - special cases have not occured yet
  - technical failures lead to new decisions
  - small decisions are made on the fly
- Documentation at the end of the project/study/creation process
  - lack of time
  - actual authors/creators maybe no longer available
  - small decisions might get lost

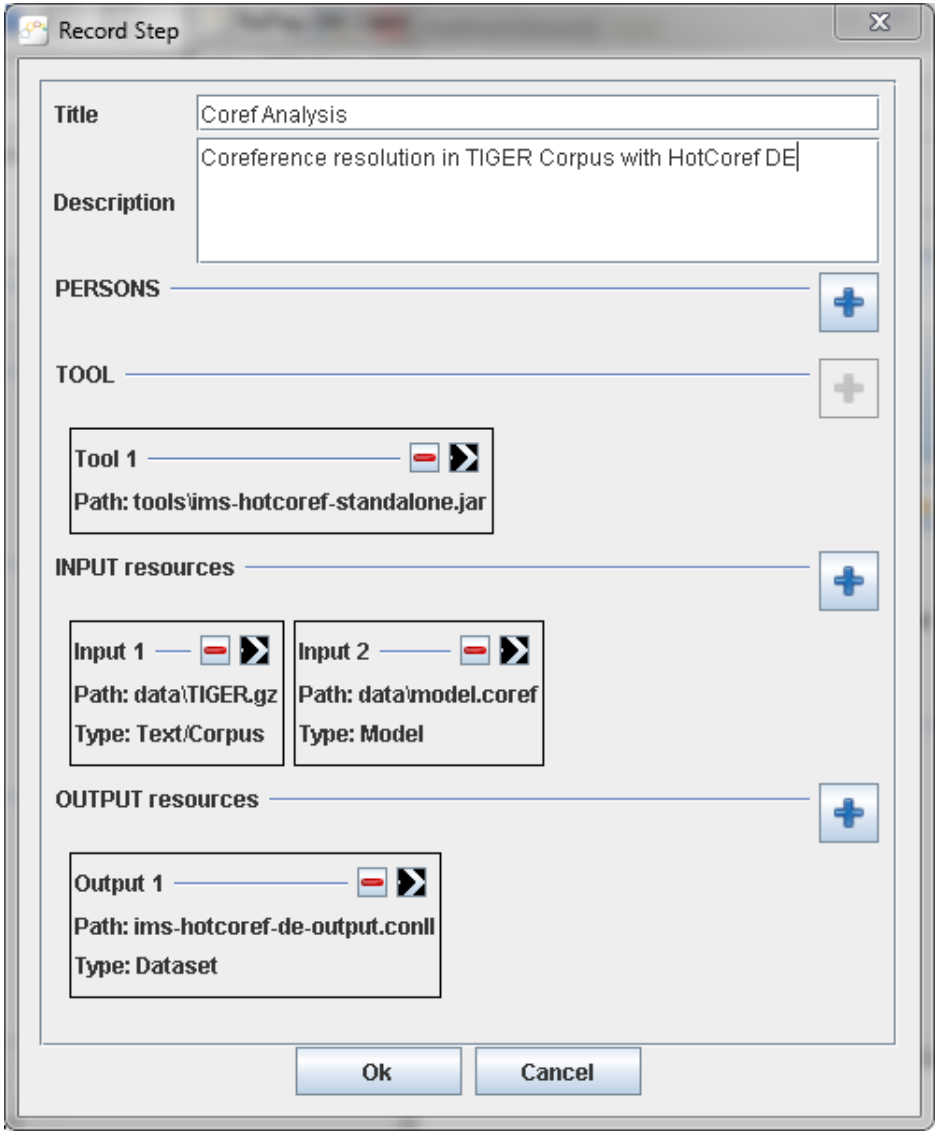
RePlay-DH Client

- Describing the workflow along the way
- Tracking fine-grained modifications
- Tracking changes and versioning via git
- Navigating through workflow history graph

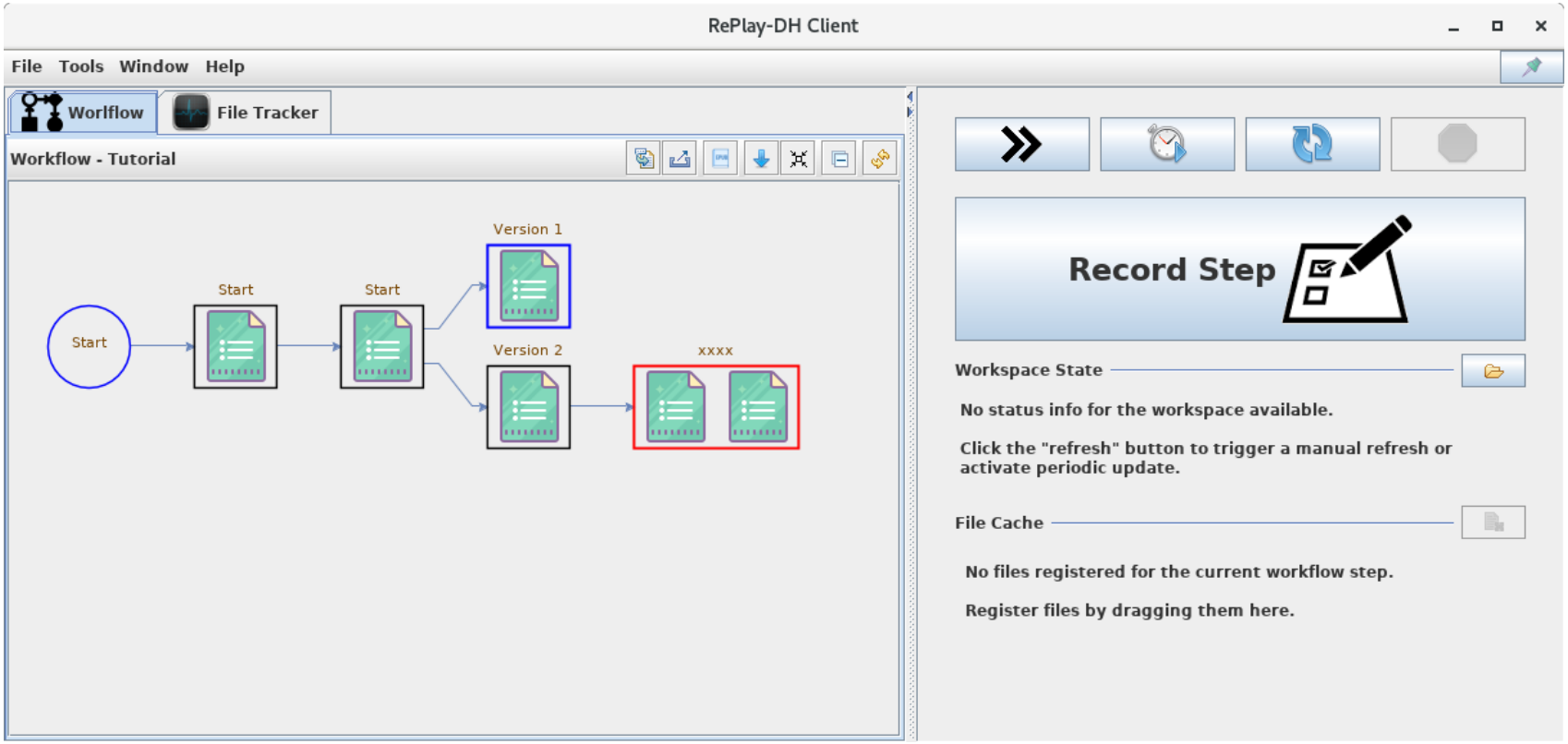
RePlay-DH Client



RePlay-DH Client



RePlay-DH Client



Process metadata …

- is an important part of the documentation of a resource/study
- makes processes transparent, fosters interdisciplinary reuse
- increases reproducibility / comparabiliy even in cases where the primary data cannot be shared
- supports the research process when tracked along the way
  $\Rightarrow$ needs to be describable with only small overhead for single researcher

Component Registry https://catalog.clarin.eu/ds/ComponentRegistry/

LAUDATIO https://www.laudatio-repository.org/

WebLicht https://weblicht.sfs.uni-tuebingen.de

[Eberle et al., 2012] Eberle, K., Eckart, K., Heid, U., and Haselbach, B. (2012). A Tool/Database Interface for Multi-Level Analyses. In *Proceedings of the eighth conference on International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Gärtner et al., 2018] Gärtner, M., Hahn, U., and Hermann, S. (2018). Supporting sustainable process documentation. In Rehm, G. and Declerck, T., editors, *Language Technologies for the Challenges of the Digital Age*, pages 284–291, Cham. Springer International Publishing.

[Gärtner et al., 2018] Gärtner, M., Hahn, U., and Hermann, S. (2018). Preserving workflow reproducibility: The RePlay-DH client as a tool for process documentation. In et al., N. C., editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[Schweitzer et al., 2018] Schweitzer, K., Eckart, K., Gärtner, M., Falenska, A., Riester, A., Roesiger, I., Schweitzer, A., Stehwien, S., and Kuhn, J. (2018). German Radio Interviews: The GRAIN Release of the SFB732 Silver Standard Collection. In et al., N. C., editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).