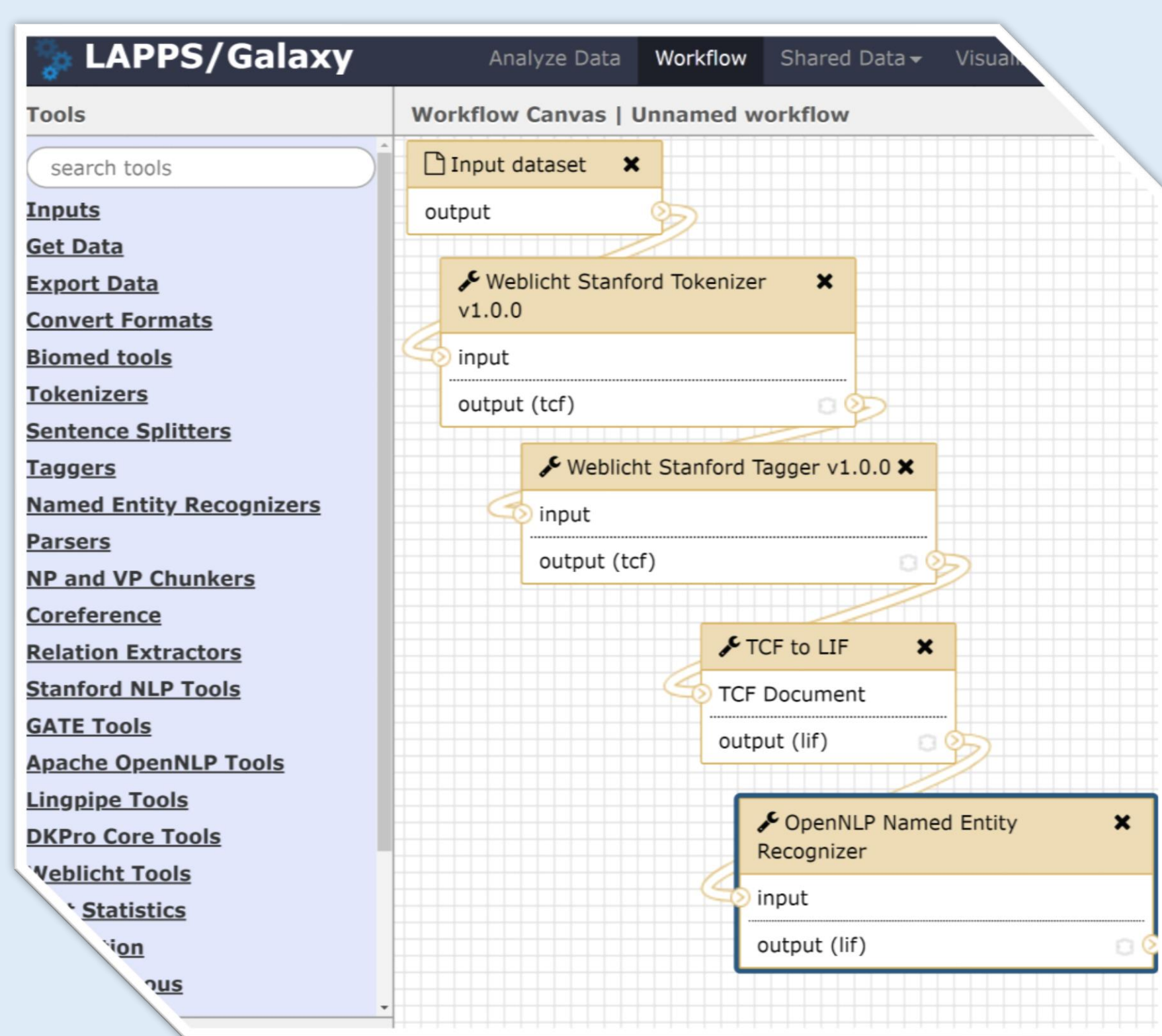# Bridging the LAPPS Grid and CLARIN

Erhard Hinrichs, Nancy Ide, James Pustejovsky, Jan Hajič, Marie Hinrichs, Mohammad Fazleh Elahi,
Keith Suderman, Marc Verhagen, Kyeongmin Rim, Pavel Straňák, Jozef Mišutka
University of Tübingen, Vassar College, Brandeis University, Charles University

THE ANDREW W. MELLON FOUNDATION

Trans-Atlantic Interoperability
Connecting Language Processing Frameworks

LAPPS Grid

- LAPPS Grid/Galaxy is used to create, execute, and visualize NLP pipelines
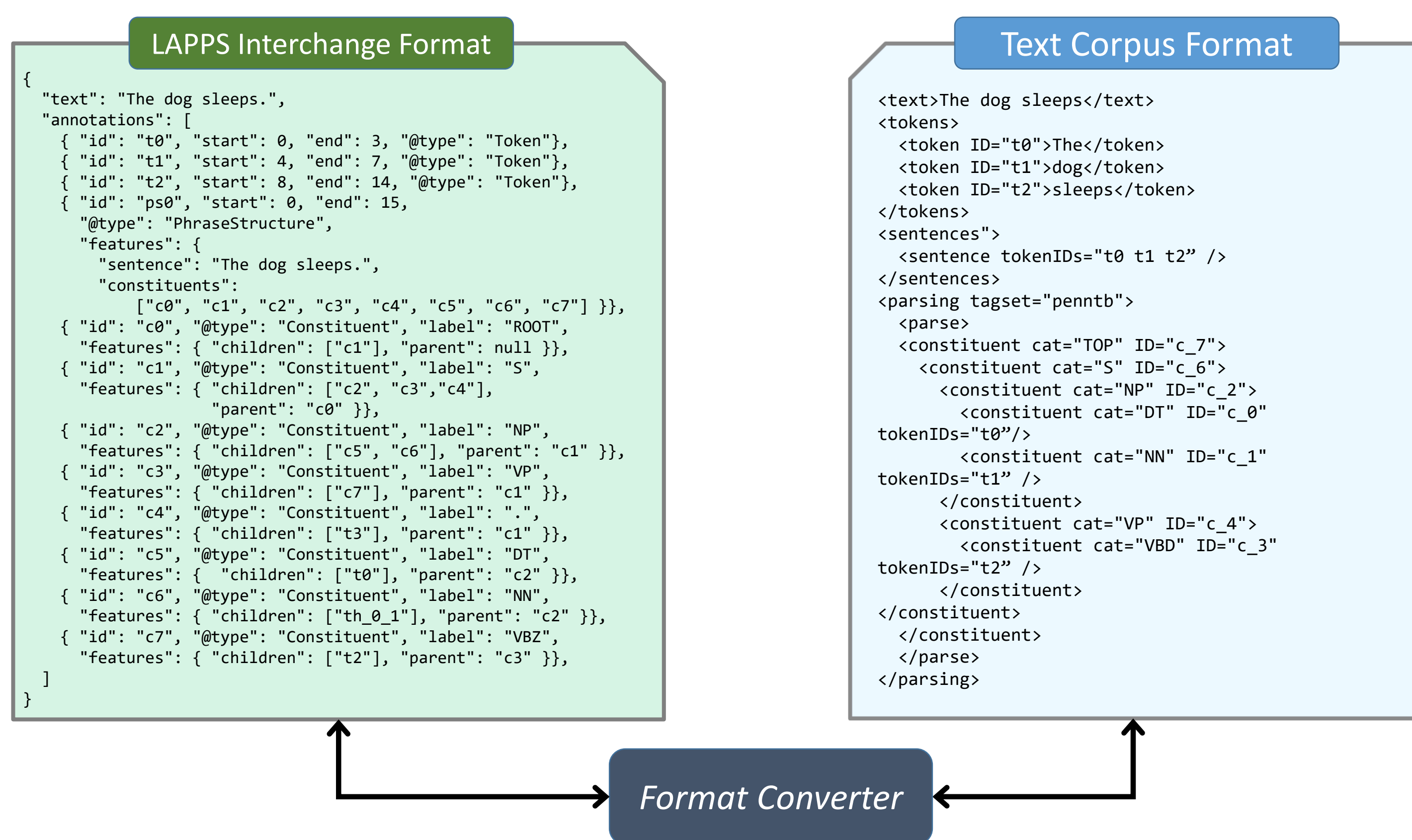- Consume and produce data in the LAPPS Interchange Format (LIF)

CLARIN/WebLicht

- WebLicht is used to create, execute, and visualize NLP pipelines
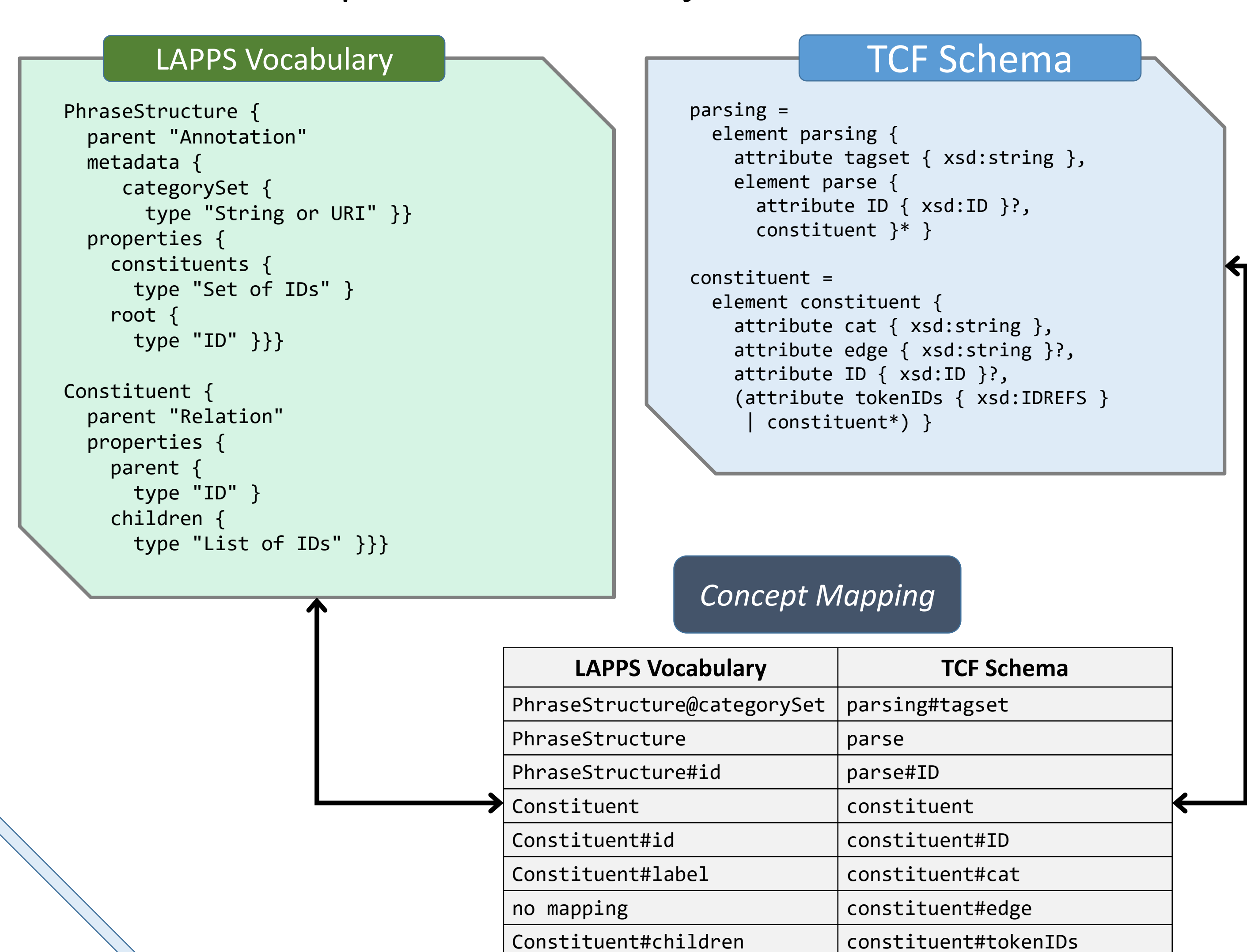- Consume and produce data in Text Corpus Format (TCF)

Syntactic Interoperability

- **LIF:**
  - flat list of JSON elements

- **TCF:**
  - XML including tag embedding
  - No text anchoring in TCF

LAPPS Interchange Format

```
{
  "text": "The dog sleeps.",
  "annotations": [
    { "id": "t0", "start": 0, "end": 3, "@type": "Token"},
    { "id": "t1", "start": 4, "end": 7, "@type": "Token"},
    { "id": "t2", "start": 8, "end": 14, "@type": "Token"},
    { "id": "ps0", "start": 0, "end": 15,
      "@type": "PhraseStructure",
      "features": {
        "sentence": "The dog sleeps.",
        "constituents":
          ["c0", "c1", "c2", "c3", "c4", "c5", "c6", "c7"] }},
    { "id": "c0", "@type": "Constituent", "label": "ROOT",
      "features": { "children": ["c1"], "parent": null }},
    { "id": "c1", "@type": "Constituent", "label": "S",
      "features": { "children": ["c2", "c3","c4"],
        "parent": "c0" }},
    { "id": "c2", "@type": "Constituent", "label": "NP",
      "features": { "children": ["c5", "c6"], "parent": "c1" }},
    { "id": "c3", "@type": "Constituent", "label": "VP",
      "features": { "children": ["c7"], "parent": "c1" }},
    { "id": "c4", "@type": "Constituent", "label": ".",
      "features": { "children": ["t3"], "parent": "c1" }},
    { "id": "c5", "@type": "Constituent", "label": "DT",
      "features": { "children": ["t0"], "parent": "c2" }},
    { "id": "c6", "@type": "Constituent", "label": "NN",
      "features": { "children": ["th_0_1"], "parent": "c2" }},
    { "id": "c7", "@type": "Constituent", "label": "VBZ",
      "features": { "children": ["t2"], "parent": "c3" }},
  ]
}
```

Text Corpus Format

```
<text>The dog sleeps</text>
<tokens>
  <token ID="t0">The</token>
  <token ID="t1">dog</token>
  <token ID="t2">sleeps</token>
</tokens>
<sentences>
  <sentence tokenIDs="t0 t1 t2" />
</sentences>
<parsing tagset="penntb">
  <parse>
    <constituent cat="TOP" ID="c_7">
      <constituent cat="S" ID="c_6">
        <constituent cat="NP" ID="c_2">
          <constituent cat="DT" ID="c_0"
  tokenIDs="t0"/>
          <constituent cat="NN" ID="c_1"
  tokenIDs="t1" />
        </constituent>
        <constituent cat="VP" ID="c_4">
          <constituent cat="VBD" ID="c_3"
  tokenIDs="t2" />
        </constituent>
      </constituent>
    </constituent>
  </parse>
</parsing>
```
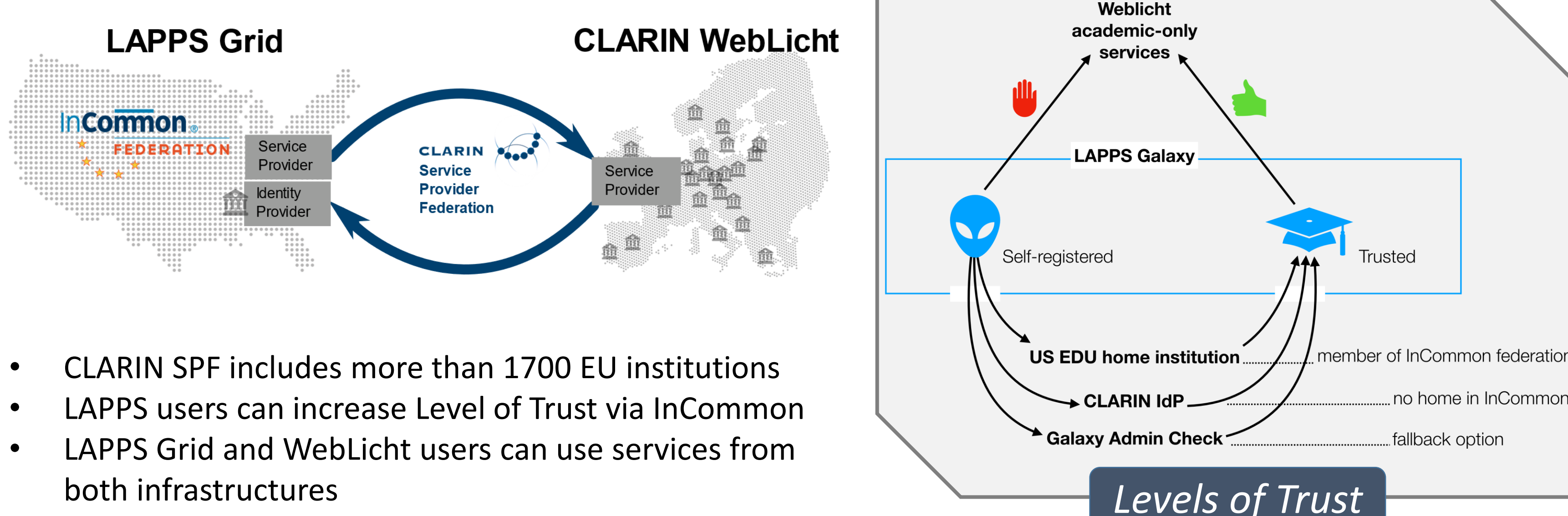
Format Converter

Semantic Interoperability

Semantic interoperability causing fewer problems for us than expected.

- Covering the core only
- No attempt to map tagsets
- Projects focus on similar tools

LAPPS Vocabulary

```
PhraseStructure {
  parent "Annotation"
  metadata {
    categorySet {
      type "String or URI" }}
  properties {
    constituents {
      type "Set of IDs" }
    root {
      type "ID" }}}

Constituent {
  parent "Relation"
  properties {
    parent {
      type "ID" }
    children {
      type "List of IDs" }}}
```

TCF Schema

```
parsing =
  element parsing {
    attribute tagset { xsd:string },
    element parse {
      attribute ID { xsd:ID }?,
      constituent }* }

constituent =
  element constituent {
    attribute cat { xsd:string },
    attribute edge { xsd:string }?,
    attribute ID { xsd:ID }?,
    (attribute tokenIDs { xsd:IDREFS }
      | constituent*) }
```

Concept Mapping

LAPPS Vocabulary	TCF Schema
PhraseStructure@categorySet	parsing#tagset
PhraseStructure	parse
PhraseStructure#id	parse#ID
Constituent	constituent
Constituent#id	constituent#ID
Constituent#label	constituent#cat
no mapping	constituent#edge
Constituent#children	constituent#tokenIDs

AAI
Authentication Authorization Infrastructure

LAPPS Grid   **CLARIN WebLicht**

- CLARIN SPF includes more than 1700 EU institutions
- LAPPS users can increase Level of Trust via InCommon
- LAPPS Grid and WebLicht users can use services from both infrastructures

Levels of Trust

Communication Protocols

LAPPS

SOAP service
Send and receive data in SOAP-XML

Web service delivers metadata **on demand** via Vassar and Brandeis servers

REST-SOAP Converter

WebLicht-LAPPS Metadata Converter

WebLicht

REST-full service
Send and receive data through direct HTTP request

Metadata follows CLARIN CMDI specification & is stored in **central repository**