Text+ : A New Research Data Initiative in Germany

CLARIN ERIC
24 January 2022
Organisers

This edition of the CLARIN Café is organized by
by Erhard Hinrichs (*University of Tübingen*), Thorsten Trippel (*IDS Mannheim*), and Andreas Witt (*IDS Mannheim*)

CLARIN hosts are
Francesca Frontini and Franciska de Jong (*CLARIN BoD*)

Technical support by
David Bordon

The event is recorded for further dissemination purposes.
Questions and comments? Put them in the chat box.
Schedule

14:00 - 14:05 - Opening - Francesca Frontini & Franciska de Jong (CLARIN ERIC BoD)

14:05 - 14:25 - Introduction to Text+ - Erhard Hinrichs

14:25 - 14:40 - Involving the Text+ Communities: SCCs & OCC - Thorsten Trippel

14:40 - 14:55 - Using a distributed technical infrastructure - Marie Hinrichs

14:55 - 15:10 - The CLARIN Standards Information Platform (SIP) - Eliza Margaretha Illig

15:10 - 15:25 - Sustainability by Forming a National Legal Entity - Andreas Witt

15:25 - 15:30 - Break for refilling coffee or tea

15:30 - 16:00 - Questions and answers
CLARIN 101

https://www.clarin.eu/content/clarin-in-a-nutshell
CLARIN ...

- is the *Common Language Resources and Technology Infrastructure*
- has the **ESFRI** ERIC status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data** (in written, spoken or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing and training**
- is an integral part of **the European Open Science Cloud**
  - See [clarin.eu/eosc](http://clarin.eu/eosc)
CLARIN today

- **70 centres**
- **22 members**: AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, SE, SI
- **3 observers**: FR, UK, ZA
The Technical Infrastructure

Harvested Metadata → Metadata Search (Virtual Language Observatory (VLO)) → Data Analysis (Language Resource Switchboard (LRS))

Indices: Findable, Accessible, Interoperable, Reusable

- clarin.eu/fair
- vlo.clarin.eu
- switchboard.clarin.eu
CLARIN and Germany - A Long Story

- in CLARIN from the start (2012, preparatory stages)
- currently 9 B-centres
- a wide range CLARIN tools and resources initiated by
  - WebLicht
  - LR SwitchBoard
The Café
Getting involved in CLARIN

Join our NewsFlash
https://www.clarin.eu/content/newsflash

Check out our events
https://www.clarin.eu/events

Open calls
https://www.clarin.eu/content/funding-opportunities
See you at the next café

... to be announced

Stay tuned: https://www.clarin.eu/content/

Share your #clarincafe impressions with @CLARINERIC
Text+:

A Sustainable Research Data Infrastructure for Language- and Text-oriented Disciplines in the Humanities

Erhard Hinrichs
Leibniz Institut für Deutsche Sprache and
Tübingen University
National Research Data Infrastructure (NFDI)

Aims of the NFDI:

» to systematically manage scientific and research data for all areas of science
» to provide long-term data storage, backup and accessibility
» to establish a network of 30 NFDI consortia for all areas of science
» to network the data both nationally and internationally.
Current State of the NFDI

» Funding of nineteen NFDI consortium and of an NFDI directorate, whose mission is to facilitate cooperation among the consortia

» Cooperation on cross-cutting topics, initially on four topics: infrastructure, training, legal and ethical issues, metadata
Two key questions:

» How many NFDI consortia are necessary for the Humanities and related fields?

» How do we arrive at a answer?

» Our solution: A workshop series in 2018, jointly organized by the Union of German Academies, CLARIN-D, and DARIAH-DE, with all stakeholders: data/infrastructure providers, universities, research institutes, research libraries, and professional associations
Complementarity and Cooperation

**Complementarity**

**Cooperation**

**Communities**
- Archaeology, Archaeometry, Archaeological Sciences, Anthropology, Ancient Architecture, Cultural Heritage Studies, Conservation Science
- History, Economic and Social History, Area Studies, Religious Studies, Historical Philosophy, Memory Institutions
- Language- and text-based disciplines, esp. General Linguistics, Modern Languages, Literary Studies, Philosophy, Classical Philology, Anthropology, non-European Cultures and Languages

**Communities**
- Architecture, Art History, Musicology, Theater, Dance and Media Studies

**Humanities**
- Data Literacy
- Rights and Ethics
- Provenance
- Metadata
- Authority Data

**NFDI**

**Text+**

For Memory

NFDI4Objects
Memorandum of Understanding as a Basis for Cooperation

## Applicant & Co-Applicant Institutions

<table>
<thead>
<tr>
<th>Co-Applicant Institutions</th>
<th>Ort</th>
<th>Co-Sprecherin oder -Sprecher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leibniz-Institut für Deutsche Sprache, Mannheim</td>
<td></td>
<td>Prof. Dr. Erhard Hinrichs (Spokesperson)</td>
</tr>
<tr>
<td>Berlin-Brandenburgische Akademie der Wissenschaften</td>
<td>Berlin</td>
<td>PD Dr. Alexander Geyken</td>
</tr>
<tr>
<td>Deutsche Nationalbibliothek</td>
<td>Leipzig + Frankfurt am Main</td>
<td>Dr. Peter Leinen</td>
</tr>
<tr>
<td>Nordrhein-Westfälische Akademie der Wissenschaften und Künste</td>
<td>Düsseldorf</td>
<td>Prof. Dr. Andreas Speer</td>
</tr>
<tr>
<td>Staats- und Universitätsbibliothek Göttingen</td>
<td>Göttingen</td>
<td>Regine Stein</td>
</tr>
</tbody>
</table>

www.text-plus.org
Additional Participating Institutions

7 Academies – 14 Universities – 3 Libraries – 5 Foundations and Centres


https://www.text-plus.org/ueber-uns/weitere-partner/
Data Domains and Thematic Clusters

COLLECTIONS
- Contemporary Language
- Historical Texts
- Unstructured Text

LEXICAL RESOURCES
- German Dictionaries in a European Context
- Born-Digital Lexical Resources
- Non-Latin Scripts

EDITIONS
- Ancient and Medieval Texts
- Early Modern, Modern, and Contemporary Texts

Erhard Hinrichs
Data Domains and Thematic Clusters

- **COLLECTIONS**
  - Contemporary Language
  - Historical Texts
  - Unstructured Text

- **LEXICAL RESOURCES**
  - German Dictionaries in a European Context
  - Born-Digital Lexical Resources
  - Non-Latin Scripts

- **EDITIONS**
  - Ancient and Medieval Texts
  - Early Modern, Modern, and Contemporary Texts

- **FURTHER DATA DOMAINS**
  - Further Clusters
Born-digitals: Authors and their Heritage

Thomas Strittmatter (1961-1995)
Drama author with a large digital legacy

» Collection:
  » digital unique material
  » digitized analog material
  » digital documents published on- and offline

» Challenges:
  » reconstruct texts from old hard- and software
  » filter legally sensitive material
Collections of Historical Texts

Degree of FAIRness

German Text Archive

Patristic Text Archive

Digital Turfan Archive

Christoph Markschies
German Text Archive
www.deutschestextarchiv.de

» active archive of 17th - 20th c. German texts with a large community
» platform for collaborative quality assurance
» standardized: via DFG-recommended DTA-Base format
» wide re-use due to analysis tools and download packages
Patristic Text Archive

https://pta.bbaw.de

» collaborative platform for critical digital editions of antique Christian texts (regardless of which ancient language)

» open access

» standardized: via a TEI-based standard

» tools for analysis
Digital Turfan Archive

» non-Western writing systems: 
~ 40,000 fragments in more than 20 different languages and scripts

» focus: Middle Iranian and Old Uyghur Buddhist, Manichaean and Christian texts (4th - 14th c.)

» Digital Turfan Archive: international cooperation including China, Great Britain, Japan, Russia
Text+ Research Data Management Strategy

FAIR and CARE

DATA SERVICES
COMMUNITY ACTIVITIES
SOFTWARE SERVICES

INFRASTRUCTURE / OPERATIONS

COLLECTIONS
- Contemporary Language – Historical Texts – Unstructured Text

LEXICAL RESOURCES
- German Dictionaries in a European Context – Born-Digital Lexical Resources – Non-Latin Scripts

EDITIONS
- Ancient and Medieval Texts – Early Modern, Modern, and Contemporary Texts

FURTHER CLUSTERS AND DATA DOMAINS

Regine Stein
Search and Interoperability Solutions

Federated data and metadata infrastructure
Search and Interoperability Solutions

Text+ schema and terminology management

GND Agency

Federated data and metadata infrastructure
Community Involvement in Text+

» calls for data, user stories, birds-of-a-feather
  → new thematic clusters, data domains, communities

» technology and methodology watch
  → new tools and software services

» training and education
  → next generation (data) humanists
Involving the Text+ Community: SCCs & OCC

Thorsten Trippel
The Community
Typical composition of infrastructure governance

If we build it, they will come.
Typical composition of infrastructure governance

If we build it, they will come.

www.text-plus.org
What if...

» ... the infrastructures are embedded in the community?
» ... the advisory board is elected from and by the community?
» ... the advisory board can influence the development by having some funds to be spent on steering the development?
» ... the infrastructures work together?
Typical composition of infrastructure governance

If they build it, they will come.
Typical composition of infrastructure governance
Text+ Governance

» Infrastructure providers: Part of the community

» Open to new partners

» Community representation
  » Functions as advisors
  » Influences developments by directly influencing funding
  » Organized in Coordination Committees per domain
  » Nominated by academic societies
  » Elected by the community of supporting societies and partners
Support by various academic societies and libraries’ specialised information services

➢ 20 learned societies and 9 specialised information services
➢ More partners to be expected


https://www.text-plus.org/ueber-uns/fachverbaende/
https://www.text-plus.org/ueber-uns/weitere-partner/

www.text-plus.org
Scientific Coordination Committee (SCC)  
Operations Coordination Committee (OCC)

The SCCs and the OCC will continuously evaluate the Text+ portfolio of data, tools, and services and will be instrumental in extending the Text+ portfolio in line with priorities set by the communities of interest. (Application, p. 34)
Task of the SCCs/of the OCC

- Participation in the allocation of Flex funds:
  - Participation in the preparation of the call
  - Review of the applications received for the data domains
  - Prioritisation of the received proposals in the SCCs as preparation for the (overarching) prioritisation in the Scientific Board
  - Review of the final reports of the projects

- Monitoring the progress of Text+
  - by means of the provided Text+ annual reports
  - through participation in the annual plenary meetings

[www.text-plus.org](http://www.text-plus.org)
Criteria for projects

Selection of extension of Text+ portfolio based on a set of common data domain and cluster specific criteria:

» Relevance and innovation
» Diversity
» FAIR and CARE principles
» Quality
» Effort

https://www.text-plus.org/forschungsdaten/user-stories/
https://www.text-plus.org/forschungsdaten/daten-aus-der-community/
https://zenodo.org/record/5384085
Community Involvement in Text+

» calls for data, user stories
» technology watch
» training and education
WebLicht
(Web-based Linguistic Chaining Tool)

Marie Hinrichs
University of Tübingen
Introduction

Overriding Goal

Pooling NLP tools and making them interoperable in an easy-to-use and open web environment

NLP tools:

- Sentence splitting
- Tokenization
- Lemmatization
- POS tagging
- Morphological analysis
- Named Entity recognition
- Dependency parsing
- Constituency parsing
• NLP tools are usually developed for a specific language

• Some of the languages supported by WebLicht:
  – English
  – German
  – Dutch
  – French
  – Italian
  – ...

• Not all annotations are available for all languages
Annotations often build upon previous annotations. Annotation tasks must be performed in a specific order.

- Parsing requires POS tags
- Named Entity recognition requires lemmas
- Sentence splitting and tokenization are required for all subsequent annotation tasks
NLP tools are typically provided as a package

- Perform all (or a subset of) annotations
- Used from the command-line or in programming code
- Examples: Stanford NLP, Open NLP, Berkeley NLP
NLP packages may have some drawbacks:

- Sometimes difficult to install
- Not always available for all operating systems
- Confusing for those not accustomed to running command-line tools or who don’t like to program
- Not possible to create some annotations with one package and other annotations with a different package
  - Input/Output formats differ
WebLicht

- Provides easy use of NLP tools online.
- NLP tools are made available by CLARIN centers as web services.
• Work started in 2009
• Makes extensive use of CLARIN infrastructure
  – Center registry
  – Center repositories
  – CMDI metadata
  – CLARIN Identity Federation
Component Interactions

CLARIN Cafe 24.01.2022

www.clarin-d.net
Easy Chains: predefined chains for the most common annotations

- Avoids having to manually build chains
- Easier for novice users
WaaS (WebLicht as a Service)

For those who prefer invoking WebLicht services from command-line, scripts, etc.

- RESTful web service to execute chains
  - Create a chain in the WebLicht interface
  - Provide WaaS with input data and chain via HTTP
WebLicht Batch (coming soon)

- Web application for WebLicht batch processing.
  - Large input files
  - Multiple input files
- Submit job for processing
- Pick up results later
Login via CLARIN Identity Federation

- Allows researchers to log in through their academic institutions
- Includes ~2200 academic institutions

https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/
Workflow

- Upload input file
- Build a tool chain
  - Easy Mode:
    - Predefined tool chains for most common annotations / languages
    - Easier for novice users
  - Advanced Mode:
    - Choose from all available tools
- Run the tools
- View / search annotations
Example

- **Input:** Wikipedia text about Douglas Adams
- **Annotations:**
  - sentences
  - tokens
  - POS tags
  - dependency parses
- **Search for all verbs in the text.**
### Example

**Search results in sentence context**

<table>
<thead>
<tr>
<th>#Sent.</th>
<th>...</th>
<th>query</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adams</td>
<td>was</td>
<td>born in Cambridge on 11 March 1952 to Christopher Douglas Adams (1927 -- 1985), a management consultant and computer salesman, former probation officer and lecturer on probationary group therapy techniques, and nurse Janet (1927 -- 2016), née Donovan.</td>
</tr>
<tr>
<td>2</td>
<td>The family</td>
<td>moved</td>
<td>a few months after his birth to the East End of London, where his sister, Susan, was born three years later.</td>
</tr>
<tr>
<td>2</td>
<td>The family moved a few months after his birth to the East End of London, where his sister, Susan, was born three years later.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>His parents divorced in 1957; Douglas, Susan and their mother</td>
<td>moved</td>
<td>then to an RSPCA animal shelter in Brentwood, Essex, run by his maternal grandparents.</td>
</tr>
<tr>
<td>4</td>
<td>A great-grandfather</td>
<td>was</td>
<td>the playwright Benjamin Franklin Wedekind.</td>
</tr>
<tr>
<td>5</td>
<td>Adams</td>
<td>attended</td>
<td>Primrose Hill Primary School in Brentwood.</td>
</tr>
<tr>
<td>6</td>
<td>At the age of nine, he</td>
<td>passed</td>
<td>the entrance exam for Brentwood School.</td>
</tr>
<tr>
<td>7</td>
<td>He</td>
<td>attended</td>
<td>the prep school from 1959 to 1964, then the main school until December 1970.</td>
</tr>
<tr>
<td>8</td>
<td>Adams</td>
<td>was</td>
<td>6 feet (1.8 m) tall by age 12, and stopped growing at 6 feet 5 inches (1.96 m).</td>
</tr>
<tr>
<td>9</td>
<td>Adams was 6 feet (1.8 m) tall by age 12, and stopped growing at 6 feet 5 inches (1.96 m).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>His form master, Frank Halford,</td>
<td>said</td>
<td>that Adams 's height had made him stand out and that he had been self-conscious about it.</td>
</tr>
<tr>
<td>10</td>
<td>His form master, Frank Halford, said that Adams had made him stand out and that he had been self-...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Statistics

<table>
<thead>
<tr>
<th>token</th>
<th>Occurrences</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>was</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>had</td>
<td>5</td>
<td>15.625</td>
</tr>
<tr>
<td>became</td>
<td>3</td>
<td>9.375</td>
</tr>
<tr>
<td>attended</td>
<td>2</td>
<td>6.25</td>
</tr>
<tr>
<td>moved</td>
<td>2</td>
<td>6.25</td>
</tr>
<tr>
<td>graduated</td>
<td>1</td>
<td>3.125</td>
</tr>
<tr>
<td>recalled</td>
<td>1</td>
<td>3.125</td>
</tr>
<tr>
<td>started</td>
<td>1</td>
<td>3.125</td>
</tr>
<tr>
<td>formed</td>
<td>1</td>
<td>3.125</td>
</tr>
<tr>
<td>wanted</td>
<td>1</td>
<td>3.125</td>
</tr>
<tr>
<td>discussed</td>
<td>1</td>
<td>3.125</td>
</tr>
</tbody>
</table>
Thank You
Eliza Margaretha Illig

CLARIN STANDARDS INFORMATION SYSTEM (SIS)

CLARIN Café on Text+: A New Research Data Initiative in Germany
24 January 2022
OUTLINE

● Background and data
● SIS tour
BACKGROUND

• SIS is a website built on XML technologies hosted at IDS Mannheim
  – XQuery, XML, XSD, HTML, JavaScript
  – eXist-db
• SIS is developed and maintained by CLARIN Standards Committee (CSC)
• SIS provides CLARIN recommendations of standards and formats for data deposition
• Github repository
  – https://github.com/clarin-eric/standards
CLARIN RECOMMENDATIONS FOR DATA-DEPOSITION

• Challenges
  - collecting standardized formats for data-deposition at CLARIN Centres
  - various ways the standards and formats are used in all kinds of Human Language Technology research
  - creating recommendation scheme
  - maintaining the recommendations
  - providing easy ways for CLARIN Centres to submit or update their recommendations

• Releases
  - https://www.clarin.eu/content/standards#clarin-recommendations-for-data-deposition-formats
# CLARIN RECOMMENDATIONS 0.1

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Mimetype</th>
<th>Extension</th>
<th>SIS ID</th>
<th>LIN</th>
<th>LINDAT</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>Wave</td>
<td>audio/x-wav</td>
<td>.wav</td>
<td>1 1 1 1 1</td>
<td>17</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Audio</td>
<td>FLAC</td>
<td>audio/flac</td>
<td>.flac</td>
<td>FLAC</td>
<td>1  1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Audio</td>
<td>ALFF</td>
<td>audio/x-aiff</td>
<td>.aiff</td>
<td>AIAFF</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Audio</td>
<td>MP3</td>
<td>audio/mpeg</td>
<td>.mp3</td>
<td>MP3</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Audio</td>
<td>RAW</td>
<td>audio/raw</td>
<td>.wav</td>
<td>RAW</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Audio</td>
<td>OGG</td>
<td>application/ogg</td>
<td>.ogg</td>
<td>OGG</td>
<td>1  1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Audio</td>
<td>BWF</td>
<td>audio/x-wav</td>
<td>.bWF</td>
<td>IBWF</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Audio</td>
<td>OPUS</td>
<td>audio/opus</td>
<td>.opus</td>
<td>OPUS</td>
<td>1  1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Audio</td>
<td>Matroska</td>
<td>audio/x-matroska</td>
<td>.mka</td>
<td>IMKA</td>
<td>1  1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Audio</td>
<td>MXF</td>
<td>application/mxf</td>
<td>.mxf</td>
<td>IMXF</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Audio</td>
<td>NIST SPHERE</td>
<td>audio/x-nist</td>
<td>.nIST</td>
<td>NIST</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Audio</td>
<td>PhonDat 1</td>
<td>??</td>
<td>??</td>
<td>IPPhonDat1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Audio</td>
<td>PhonDat 2</td>
<td>??</td>
<td>??</td>
<td>IPPhonDat2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Compression</td>
<td>GZIP</td>
<td>application/gzip</td>
<td>.gz</td>
<td>I1GZIP</td>
<td>1  1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Compression</td>
<td>ZIP</td>
<td>application/x-zip</td>
<td>.zip</td>
<td>I1ZIP</td>
<td>1  1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Compression</td>
<td>RAR</td>
<td>application/x-rar</td>
<td>.rar</td>
<td>I1RAR</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Computer Aided Design</td>
<td>DXF</td>
<td>application/dxf</td>
<td>.dxf</td>
<td>DXF</td>
<td>1  1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Computer Aided Design</td>
<td>SVG</td>
<td>image/svg+xml</td>
<td>.svg</td>
<td>ISVG</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Computer Aided Design</td>
<td>REFl/QDA</td>
<td>application/x-refl</td>
<td>.QDA</td>
<td>REFl/QDA</td>
<td>1  1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Container</td>
<td>TAR</td>
<td>application/x-tar</td>
<td>.tar</td>
<td>ITAR</td>
<td>1  1</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
RECENT CLARIN RECOMMENDATIONS

• Complete recommendations

--<recommendation>
  --<header>
    --<title>CLARIN Standards Information System (SIS) export</title>
  --<url>
    https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq
  --<date>2022-01-31T16:21:24.255+01:00</date>
  --<filter/>
  --</header>
  --<formats>
    --<format id="TAAC">
      --<centre>DANS</centre>
      --<domain>Audiovisual Source Language Data</domain>
      --<level>acceptable</level>
    --</format>
    --<format id="TADS">
      --<centre>BAS</centre>
      --<domain>Audiovisual Annotation</domain>
      --<level>recommended</level>
    --</format>
  --<format id="TALK">
    --<centre>DANS</centre>
    --<domain>Image Source Language Data</domain>
    --<level>acceptable</level>
  --</format>
  --<format id="TAPI">
    --<centre>IDS</centre>
    --<domain>Audiovisual Source Language Data</domain>
    --<level>acceptable</level>
  --</format>
--</recommendation>

• Centre recommendations

--<recommendation>
  --<header>
    --<title>CLARIN Standards Information System (SIS) export</title>
  --<url>
    https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq
  --<date>2022-01-31T16:23:32.046+01:00</date>
  --<filter/>
  --<centre>IDS</centre>
  --<header>
  --<formats>
    --<format id="TIAFF">
      --<domain>Audiovisual Source Language Data</domain>
      --<level> acceptable </level>
    --</format>
    --<format id="TALTO">
      --<domain>Text Annotation</domain>
      --<level> acceptable </level>
    --</format>
    --<format id="TANVIL">
      --<domain>Audiovisual Annotation</domain>
      --<level> acceptable </level>
    --</format>
    --<format id="TCHAT">
      --<domain>Audiovisual Annotation</domain>
      --<level> deprecated </level>
    --</format>
--</recommendation>
SIS TOUR

https://clarin.ids-mannheim.de/standards/
THANK YOU!
CLARIN Café on Text+: A New Research Data Initiative in Germany
24 January 2022

Andreas Witt, Department Digital Linguistics at IDS, CLARIN ERIC

SUSTAINABILITY BY FORMING A NATIONAL LEGAL ENTITY
SUSTAINABILITY ISSUES OF DIGITAL RESEARCH INFRASTRUCTURES.

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Topics</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical</td>
<td>FAIR, data curation, meta data, standards...</td>
<td>Partners and infrastructures</td>
</tr>
<tr>
<td>Legal</td>
<td>IP licenses, GDPR-compliance, contracts...</td>
<td>Expertise, professionalization</td>
</tr>
<tr>
<td>Financial</td>
<td>Short-term vs. long-term, openness...</td>
<td>Operating model</td>
</tr>
<tr>
<td>Organizational</td>
<td>Processes, community, demands, usability...</td>
<td>Legal form</td>
</tr>
</tbody>
</table>
CLARIN IN GERMANY.

Funding on project basis: Building a national community from the bottom up.
- 2008–2011 as D-SPIN.
- 2011-2019 as CLARIN-D.
- 2019-2021 CLARIAH-DE: Synergy between the major infrastructures of the German humanities.

Financing at European and national levels.
- Annual membership fee to the European research infrastructure consortium CLARIN.
- Providing a national organization of the German CLARIN community.

National Perspectives.
- No public funding basis after finalization of the national projects.
- Service operation at the expense and expertise of the national CLARIN centers.
- Demand for an organizational form to coordinate providers, users and services.
- Keep in touch with the European activities and with other research data activities in Germany (NFDI).
## THE STATUS SO FAR.

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Topics</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical</td>
<td>Migration of CLARIN-D services.</td>
<td>Text+</td>
</tr>
<tr>
<td>Legal</td>
<td>Migration of CLARIN-D expertise.</td>
<td>Text+</td>
</tr>
<tr>
<td>Financial</td>
<td>Migration of CLARIN-D expertise.</td>
<td>Membership in CLARIN ERIC</td>
</tr>
<tr>
<td>Organizational</td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>
• Text+ is a large community
  • CLARIN, DARIAH, German National Library...
• Text+ is part of the German National Research Data Infrastructure (NFDI)
  • Task: build a research infrastructure for text- and language-based data
• Text+ and CLARIN
  • Text+ ≠ CLARIN-D
  • Many CLARIN-D centres are Text+ members
  • CLARIN-D has an independent National Coordinator
TEXT+: WHAT IT NEEDS TO THRIVE.

- **Text+ needs a channel to exchange with CLARIN**
  - CLARIN works (e.g. standards, centre certification...)
  - CLARIN's experience could provide major insights for Text+
  - CLARIN community has specific needs that Text+ should address

- **Solution:**
  - A CLARIN community on a national level (CLARIN centres etc.)
  - A legal entity
IN SEARCH OF A LEGAL FORM.

What we need.
Community requirements for a dedicated legal form.
- Tried and tested solution.
- Non-profit status.
- Institutional beneficiary of public funds.
- Low or no founding and operating costs.
- Openness towards new members.
- Democratic structure.
- Strong community participation.
- Modular membership.
What we need: an association („Verein“)
What we did: We transformed and renamed the association TextGrid e.V. (est. 2013)
Why?:
- to interlink the CLARIN-community.
- to get a humanities ambassador to relevant ERICs or other consortia.
- to establish a contact point for researchers, stakeholders, infrastructures, funders.
- to benefit from the huge overlap of former TextGrid members and CLARIAH partners.
- to sustainably provide CLARIN services collaboratively.
- to shape future services and dynamics at the national level.
- to rely on a proven non-profit legal form without much red tape.
WHO ARE THE MEMBERS?

26 universities, academies, extramural research institutions, service providers, cultural heritage organizations, archives, libraries and individuals...
HOW IT WORKS.

**MEMBERSHIP**
- legal entities
- full vs. associate vs. ex officio members
- modular participation opportunities
- resource contributions vs. expertise and counselling
- full beneficiary of service portfolio

**STRUCTURE**
- general assembly: grand strategy, elections, funds.
- coordinating council: service portfolio, interface to the world
- working groups: well, real work...

**SERVICES**
- research
- access to funding
- counselling
- resource provision
- advancement
- networking

**COOPERATION**
- national humanities
- NFDI, national infrastructures
- ERICS
- funders
- government
- multipliers

Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V.