CLARIAH Linguistics Plan.

Jan Odijk, Daan Broeder & Sjef Barbiers

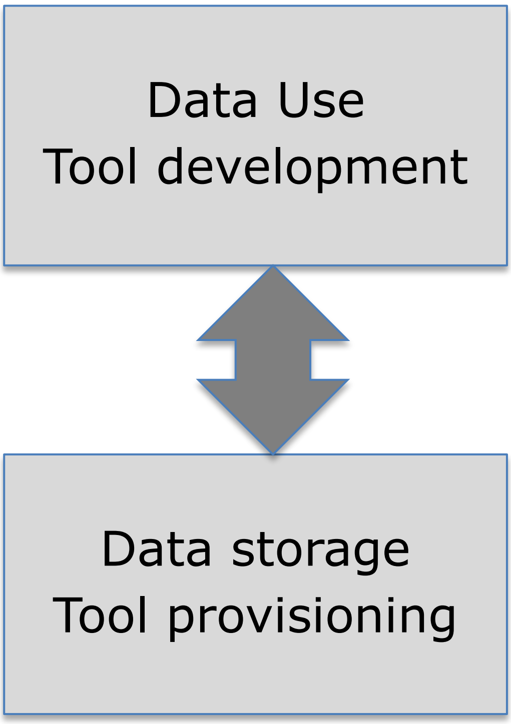
Version 0.95

Contents

1. Introduction

CLARIAH is a research infrastructure that aims to support research by humanities researchers. It must therefore provide facilities for the workflow in a typical research project that provides & creates its own data next to enabling access to the large standard data sets that have been gathered over the years. In such a project, the focus is on research, and data and tools are just instrumental.[1] Data created in a typical research project are relatively small in size (compared to e.g. Nederlab). Exact maximum size of external or newly created data for which CLARIAH wants to support their incorporation in the infrastructure still has to be determined, but it will be orders of magnitude lower than the Nederlab data. For incorporation of data of the size of the Nederlab data a dedicated project is required, and probably dedicated software has to be developed.

The strategy chosen for the CLARIAH Linguistic RI is that we provide long-term data availability together with high quality metadata for a wide range of usages. While for the data consumer side (research) we provide flexible frameworks populated with tools and usage recipes directed at specific research use-cases.



- Research specific purposes
- Flexibility
- Short term lifecycles

- General services
- Long term lifecycle
- Long term data persistency
- Research verfiability
- Data policy enforcement

CLARIAH has to offer functionality that a researcher needs. This includes:

- Obtaining Data
    - Search for and select data already contained in the CLARIAH infrastructure

[1] This is completely different from the perspective of infrastructure projects themselves such as CLARIN-NL or Nederlab, which are research infrastructure projects in which the focus is primarily on the data and the tools.

  - Incorporate existing data into the CLARIAH infrastructure
  - Create new data, inter alia via crowd sourcing, and incorporate them into CLARIAH
- Obtaining tools (software services)
  - Search for and select tools that are already part of the CLARIAH infrastructure
  - Incorporate existing tools into the CLARIAH infrastructure
  - Create new tools, and incorporate them into CLARIAH
- Enriching data incorporated in CLARIAH with various annotations
  - Automatically
  - Manually, possible automatically bootstrapped
  - Linking to external resources
- Searching in and analysis of the data
  - Upload (possibly enriched) data into a search engine
  - Search and browse in the data, analyze the data and the results of searching
- Visualising search and analysis results
- Publishing the data and software in the CLARIAH infrastructure
  - Make them visible (through metadata), accessible, and referable
  - Ensure safe long term storage
- Creating and publishing enhanced publications (scientific article plus associated data and tools)

This plan describes each of these items in more detail. Since the CLARIAH infrastructure is a distributed infrastructure, each element of this functionality must be provided through at least one of the CLARIAH centres or other approved provisioning organizations.

We relate these elements to the research work flow as sketched by Hennie Brugman (which itself is based on a work flow in use within Nederlab) in an appendix (see section 5).

We refer to the thematic plans submitted by the theme leaders and to the CLARIAH Linguistics projects Excel workbook [Odijk 2014c] wherever appropriate. We identify tasks to be carried out and assign them an identifier (of the form [Tii] with *i a* digit*)* for ease of reference in external documents (e.g. the accompanying task descriptions in Excel).

[T00] The overall architecture of the infrastructure, covering all items mentioned above, and ensuring their interoperability, must be designed and implemented [Odijk 2014c tasks 26]/[Brugman 2015 Task 1.1] are part of this.

Some functionality is required in virtually every stage of the research workflow. This concerns in particular *syntactic and semantic interoperability*. We have described these aspects in one of the sections where it, in our view, fits best: section 2.1.2.

We assume that **[T01]** Nederlab will form an integrated part of CLARIAH and some work may be required to fully integrate it (since CLARIAH might have slightly different requirements) [Odijk 2014c Task 18]

For each item listed above we will describe

- What it should ideally contain
- What we already have (from CLARIN-NL and/or other projects)
- What still needs to be created anew. For new tools that require cutting-edge or experimental technology it is strongly advised to first carry out an assessment of their feasibility (e.g. is the state of the art sufficiently advanced, are the necessary data for training in place, is there enough knowledge and expertise ,etc. or a proof-of-concept project before investing significantly.
- What of the existing data and tools has to be updated or upgraded
- Parties that should be involved

- References to thematic work plans and other documentation.

Overall we assume that research is conducted in a research group and that, by default, all members of the research groups (and only these) see each other's data and that further sharing takes place by publication through a CLARIAH center's repository. Though a Virtual Research Environment (VRE) for this purpose would be desirable, the functionality of such a VRE is still badly defined and several aspects of it can be dealt with by existing software (e.g. Dropbox or Surfdrive for a shared workspace). We assume that the portal to be created in WP1 can be used to guide the user through the various options offered in different stages of the research. Other facilities will be part of individual applications and services.

We have to make concrete arrangements for dealing with version control and configuration management inside the CLARIAH project, both for the process before publication of data and tools, and for the versioning of published data and tools. This has to be done within WP3 but in close cooperation with the other WPs, in particular WP2.

We assume that a researcher either brings his/her own data and tools, which have to be brought into the CLARIAH infrastructure, or selects data/tools that are already included in the CLARIAH infrastructure. We also assume that inside the CLARIAH infrastructure, only a few explicitly chosen data formats for data will be supported, and that all data have associated (CMDI) metadata. This holds not only for textual data, but also for audio-visual data and for structured data. We will call these *the CLARIAH-internal data formats*. For new data brought into the infrastructure, facilities must be offered to convert the data of a researcher into one of the CLARIAH-internal data formats. Tasks for this are defined in section 2.1.2.

It is important that the quality of data, tools and their metadata provided by the CLARIAH infrastructure is of high level, and that their metadata contain formalized information that is crucial for discovering resources. Therefore special facilities will be provided for curation of data and metadata. The results of this curation process will be made available to the (meta-) data providers for republishing, but also for immediately improving the quality of tools such as the CLARIN metadata catalogues (e.g. VLO) and (aggregated) content search tools.

2. Work Flow Items

2.1 Obtaining Data

2.1.1 Data Selection

[should contain] Tools to browse and search for data inside the CLARIAH/CLARIN infrastructure that a researcher might want to use in his/her research, and options to select subsets of these data (virtual collection) and retain these selections for a longer period of time. The searching and browsing is possible thanks to metadata, which for WP3 will be at least in CMDI-format, and will also be made available as Linked Data (LD), in a format agreed upon with WP2).

[what we have] VLO, Meertens Metadata Search, CLARIN-NL Portal faceted search for Data. Metadata for many data; these metadata are harvestable via OAI-PMH at Meertens and INL, and they are currently harvested by the maintainers of the VLO. System for registering virtual collections (i.e. extensional data selections associated with a PID and metadata) from CLARIN EU. CMDI2RDF module based on Virtuoso Store.

[what is newly needed] The VLO has become difficult to use by its success: it contains so many data from so many different origins that it has lost its function to find data [Odijk 2014b], and it requires significant adaptations in the VLO and in the metadata to regain this function. **[T02]** We should experiment with a NL only VLO (or the Meertens Metadata Search) to set a good example for the redesign of the CLARIN-wide VLO. One CLARIAH-centre should host this and do metadata harvesting within NL. **[T03]** We also should curate the CMDI metadata to have them contain the obligatory metadata info (we discuss this separately in section 2.1.4). We need at least the possibility to filter on collection type of metadata only. The CLARIN portal data, exported and converted to CMDI, might form a good basis for this.

Investigating other possibilities of mapping CMDI to LD, e.g. on the basis of XSLT, and from LD to CMDI will be carried out in WP2.

[Parties] Meertens and INL. UU.

[References] [Brugman 2015 Task 1.3]; [Odijk 2014b]; [Odijk 2014c task 24]

2.1.2 Incorporate External Existing Data

[should contain] Facilities to incorporate an existing dataset that is not yet part of the CLARIAH infrastructure into the CLARIAH infrastructure. This requires tools to convert actually used formats into the CLARIAH internal data formats, and tools to make the semantics of the data explicit. It also requires tools to create metadata for the data. A service to develop such tools, carry out adaptations (which form a major part of what is called *data curation*) for selected resources, and assist researchers in curating their data, is needed.

[what we have]

- Data Curation: Experience at the CLARIN-NL Data Curation Service (DCS).
- Data Conversion: Some tools are currently available or are being developed by INL in the CLARIN-NL OpenConvert project (.doc ->HTML, ePub->HTML, (x)HTML->TEI; docx-> TEI; ALTO->TEI; FoLIA->TEI; TEI->FoLIA; ALTO->FoLiA; Word -> plain text; ePub -> plain text); Use can be made of pandoc to extend this set.
- PICCL / TICCL
- CLARIN Concept Registry (CCR) for specifying the semantics of data and metadata elements.
- Tools to make CMDI metadata profiles and components and store them (CMDI Registry), and a variety of CMDI metadata editors (Arbil, COMEDI, …)

[what is newly needed]

- **[T06]** Determination of the *CLARIAH internal data formats* [Vossen 2015 task 2.1, though broader than described there], starting from the list of standards provided by CLARIN. It is likely that **[T70]** FoLiA will be one of these formats, and support for this format should be provided [Van den Bosch 2015 Task 2.3]
- **[T07]** Tools to convert often occurring formats used by linguists into the formats supported inside the CLARIAH infrastructure, not covered yet by what we already have. See [Odijk 2014a] for an overview.
- **[T08]** Facilities to easily create CMDI metadata for often recurring data types (lexicons, text corpora, audio corpora), at least for the collection level, with all required metadata information and with as little requirements for knowledge/expertise with CMDI.
- **[T10]** Data Curation Service [Van den Bosch Task 3.2, 3.4]. The list created by the CLARIN-NL DCS, as well as known needs from running or starting research projects can serve as a good basis for the selection of the data to be curated.
- **[T11]** Facilities to make the semantics of data and metadata elements explicit by linking them to data category and concept registries (CCR, ISO Language Codes, etc.) via CLAVAS. [Vossen 2015 Task 2.2, 2.3, 2.4] [Odijk 2014c Task 29]
- **[T12]** Add relations to the elements in the CCR and **[T13]** use these relations in search (more precise/ less precise search) and also **[T14]** in the CCR interface to group closely related concepts [Odijk 2014c Task 30]

[updates/upgrades]

- **[T15]** CMDI registry must facilitate easier searching of relevant CLARIAH recommended profiles and components

[Parties] Meertens, RUN, INL for their converters and extensions of them; Semantic interoperability: VU, UU (Ineke Schuurman), Meertens (CCR, CLAVAS)

[References] [Brugman 2015, Tasks 1.1-1.3; [van den Bosch 2015, task 3.1, 3.2, 3.4]; [Odijk 2014a] [Odijk 2014c task 14, 29, 30]

2.1.3 Create new data and incorporate them into CLARIAH

[should contain]

1. facilities to collect new data, inter alia via crowd-sourcing, and incorporate the resulting data into the CLARIAH infrastructure. This includes not only textual data but also structured data and audio(-visual) data
2. facilities to incorporate pictures of text enriched with textual transcriptions via PICCL
3. facilities to extend databases (e.g. typological databases) with new data

[what we have] Some crowd-sourcing and survey tools at Meertens; probably open-source tools that can be used/adapted for crowd-sourcing and surveys via the web. PICCL. TDS (typological database system) exists but the status of its editing and uploading functionality has to be checked and possibly extended.

[what is newly needed] [T09] Inventory of existing tools for crowd-sourcing and surveys. Selection of preferred tool(s). New crowd-sourcing and survey software or adaptation of existing software (not covered yet in any of the themes). WP2 includes SurveyMonkey in its task 42.100, but WP3 will work on crowd-sourcing, esp. for transcription, and formulate requirements and desiderata for surveying software.

[updates/upgrades] [T09] Existing tools may have to be tuned to the specifics of data collection by linguists. Documentation targeted at linguists and using linguistic examples may be required. All selected crowd-sourcing and survey tools should be adapted to directly yield data in CLARIAH-internal formats and with CLARIAH-compatible metadata. An update/upgrade of TICCL is needed but is relevant for multiple disciplines in the humanities, so will be done in WP2 (41.400). **[T26]** Upgrade of PICCL [van den Bosch 2015 Tasks 1.3, 3.3]. **[T18]** Extend TDS with editing and upload functionality

[Parties] Meertens, UvT (PICCL), UU/Meertens (TDS)

[References] [van den Bosch 2015 Task 3.3] [Odijk 2014c Task 28]

2.1.4 Metadata Curation

[should contain] There were hardly any requirements concerning the content of metadata for the data and software created in CLARIN-NL. Though flexibility is required, and we want the owner/researcher of a data set to have a big say in the selection of metadata elements for a data/software set, we also have to consider the overall context in which these metadata will end up: otherwise it will be impossible to find any data. See [Odijk 2014b] for clear examples of this problem. Within CLARIAH we can impose requirements on metadata on the partners financed by CLARIAH, e.g. requiring the use of specific profiles, components and/or metadata elements (very likely slightly different per research discipline). In addition, we should start up a discussion with the CLARIN CMDI task force so that such a limited set of recommended/required CMDI profiles, components, or metadata elements gets wider usage and can be specifically exploited in metadata search and browse engines such as the Virtual Language Observatory (VLO). We also need ways to improve errors in the current available CMDI metadata. The most efficient way is for a dedicated team to monitor the harvested metadata and repair and/or map to a new set of recommended CMDI profiles.

[what we have] The ideas from the CLARIN CMDI task force and [Odijk 2014b]

[what is newly needed] [T03] We need to determine the minimal set of required metadata elements (or components) for the metadata (this may be discipline-specific). We must start up an action to upgrade all existing CLARIN-NL metadata in this way. ([Brugman 2015 task 1.3]). A feedback loop of the curated improved metadata to the originator has to be created. **[T04]** We need CMDI record evaluation tool(s) that checks compliance with the requirement for occurrence of specific metadata elements or components, or checks for the use of specific profiles.

[Parties] Meertens, INL and RU (DCS). UU will be involved for specifying requirements on metadata.

[References] [Brugman 2015, Tasks 1.1-1.3; [van den Bosch 2015, task 3.4]; [Odijk 2014a] [Odijk 2014c Task 25]

2.2 Obtaining Tools

2.2.1 Search for and Select Tools existing in the CLARIAH infrastructure

[should contain] Facilities to browse and search for software through descriptions of this software (their 'metadata'). Searching should be possible as much as possible via facets with a closed or half-open value set.

[what we have] VLO. The CLARIN-NL Portal Services part; CMDI profile for software

[what is newly needed] [T19] Metadata for the existing tools must be created. For the discovery part of such metadata, the information of CLARIN-NL Portal forms a good basis. The technical part still has to be extended by the software developers and will very likely still have to be fine-tuned.

[updates/upgrades] The VLO supports this in principle, but currently (1) almost none of the Dutch tools have metadata; (2) the VLO offers no proper facets for searching for metadata. **[T02]** We could extend the local copy of the VLO to support faceted search for software as well. To facilitate easy curation of tool metadata, the CLARIN Portal tool and services descriptions can be transformed into CMDI and be made OAI-PMH harvestable.

[Parties] Meertens, INL, UU (JO), all software developers will have to be provide some input in their data,

[References] none

2.2.2 Incorporate existing tools into CLARIAH

[should contain] Specifications of requirements CLARIAH-compatible software must meet. Facilities to ensure that existing software can be easily made CLARIAH-compatible. Facilities to make CLARIAH-compatible metadata for the software.

[what we have] very little, we think. CLAM is a clear example; the CMDI metadata profile for software.

[what is newly needed] We will take into account guidelines issued from WP2 (task 54.700), and, if needed, make additional arrangements within WP3.

[updates/upgrades] [T21] CLAM must continue to be maintained and supported [Van den Bosch 2015 Task 2.4]

[Parties] RUN, Meertens, UU

[References] [Van den Bosch 2015 Task 2.4]

2.2.3 Create new tools and incorporate them into CLARIAH

[should contain] Specifications of requirements that CLARIAH-compatible software must meet. Recommendations for using specific APIs, web services, protocols, etc. Facilities to make CLARIAH-compatible metadata for the software.

[what we have] Some guidelines w.r.t. federated login, web service protocols, CLARIN open software creation guidelines.

[what is newly needed]

We will take into account guidelines issued from WP2 (task 54.700), and, if needed, make additional arrangements within WP3.

[updates/upgrades] none

[Parties] Meertens, RUN

[References] none

2.3 Enrichment and Annotation

[should contain] A wide range of software which enables a linguistic researcher to enrich his/her data with all kinds of (mainly linguistic) annotations, among them annotations to link data to external resources (e.g. a linguistic expression to an entry in a database). The software should include tools to carry out automatic enrichment but also tools to do manual addition, verification and/or correction of (automatically enriched) data. Ideally these tools take as input not only data files but also their associated (CMDI) metadata files and yield as output new, updated data files with their associated metadata files (incl. provenance information). The output metadata files should contain information about all parameters settings and/or a reference to a configuration file with these parameter settings that can be used as input to the (automatic enrichment) tools so that replication of the action is made easier.

We may and must require that the input formats of the data are restricted to only a few CLARIAH-internal data formats (see section 2.1.2), and that other formats have already been converted into one of the CLARIAH-internal data formats.

The output formats should be such that Search and Analysis tools from the box **Search & Research** can accept them as input to make them searchable and analyzable.

We need a facility to mark and administrate the resulting new (enriched) annotations in such a way that the relation to the original primary data and/or annotations is maintained. If the primary data is hosted at a CLARIAH center and when agreed by the maintainer of the primary data collection, it should be possible to deposit the new data alongside the original. It is clear that is foremost interesting for manual annotations (but also or automatically generated annotations that require large computational resources).

[what we have] for automatic enrichment: A range of tools exist, among them TTNWW and all tools contained in it; Frog at the Nijmegen server (currently not part of the CLARIN infrastructure), Alpino parsing inside PaQu; PICCL; Adelheid & INPOLDER; Namescape; NERD; Software designed to support the manual annotation process such as AAM-LR and TQE.

For manual annotation, FLAT (for annotation of FoliA-format data) is under development. ELAN/ANNEX for A-V-data; Praat (not currently part of the CLARIN infrastructure). Opensource tools such as TrED (Treebank editing), Annis, and undoubtedly many others may be considered

[what is newly needed] There is a desire to obtain **[T78]** (better) software for the linguistic enrichment (morphosyntax, morphology, syntax) of older variants of Dutch. **[part of T22]** Experiments could be done with Alpino and Frog by using a `analyze as' feature (analyze word *a* as if it were word *b*). **[part of T23]** Or develop a new Frog for older variants of Dutch based on a set of training data from this variant (provided such a set exists and/or can be made with limited effort) [Van den Bosch 2015 task 2.2], [Odijk 2014c Task 21].

[updates/upgrades]. It is likely that Folia will be one of the CLARIAH-accepted internal formats, so **[T24]** further development of FLAT is required [Van den Bosch Task 2.5]. However, making a good and workable annotation tool is an art in itself, and the adequacy of a tool is often inversely proportional to all the options it has. Furthermore FLAT is a web-based linguistic annotation tool, and so far one of us (JO) has never seen a decent web-based interface for anything (except Google Docs and MS Outlook webmail) and certainly not for annotation, so we should not put all our horses on FLAT, but provide support for other selected annotation tools as well, even for textual resources. The use of local annotation tools can be necessary both from the usability and the network independence side. **[T22]** Upgrading Frog [Van den Bosch 2015 Task 2.1]. **[T26]** Updating PICCL. Updating TICCL [Van den Bosch 2015 Task 1.3] is covered in WP2 (see section 2.1.2). **[T55]** improvement of UCTO [Van den Bosch Task 1.1].

The basic functionality of TTNWW (i.e. automatic enrichment by uploading a file and pushing a button for executing a preconfigured recipe of actions) must be retained, though we should investigate more efficient ways of service deployment **[T28]** However, its interface (both for input and for output) has to be improved significantly, the output formats must be in one of the supported CLARIAH-internal data formats (currently there are many ad-hoc formats), possibilities to upload a zipped file and/or a whole directory or a selection of files based on pattern must be added. It should also accommodate other input formats than just plain text, or else provide for automatic conversion, though this can be restricted to the CLARIAH-internal data formats (see section 2.1.2).

TTNWW currently contains older versions of software (e.g. TICCL and Alpino) than are used and published by the developers themselves: **[T28]** TTNWW should be upgraded with these new versions and a strategy should be found to deal with such future updates and upgrades in as easy a manner as possible. See [Odijk 2014a].

Close attention must be paid to work with the data & metadata infrastructure (theme 1 and WP2 facilities) for keeping track of new versions of annotations.

[Parties] Nijmegen, Meertens (for TTNWW)

[References] [Odijk 2014a] [Odijk 2014c Tasks 21, 22, 23]; [Van den Bosch 2015 task 1.1, 1.2,1.3, 2.1,2.2]

2.4 Searching and Analysis

2.4.1 Upload Data into a Search Engine

[should contain] There must be facilities for a researcher to upload a corpus that has been enriched in a CLARIAH-compatible manner (e.g. with CLARIAH enrichment tools) to a search engine and to make it searchable in this way through an existing search engine and application.

[what we have] Upload and Search: Initial versions of corpus upload and search functionality have been created in the CLARIN-NL PaQu and AutoSearch projects

[what is newly needed] [T31] Upload functionality should be added for GrETEL

[updates/upgrades] The functionality of **[T32]** PaQU and **[T33]** Autosearch should be extended (especially support for more input formats and **[T49]** upload of data AND metadata. Initial set of desiderata is available.

[Parties] INL (AutoSearch), RUG (PaQu), UU (upload functionality for GrETEL)

[References] [Odijk 2014c Task 2]

2.4.2 Search and Browse

[should contain] Facilities to browse and search in data (content) and their metadata and to carry out analysis on the search results by grouping, sorting, filtering and combining these results. Ideally one query need to be launched to search in several data sets *of the same type* even if these are distributed over multiple locations (restricted federated content search) or by consolidating the annotations under a single search engine by harvesting the annotations from different sites. A user must be able to select the data he/she wants to browse and search in, e.g. via a browse and search engine for metadata from which the relevant data can be selected on the basis of their metadata.

The concept *of the same type* still has to be defined in more detail, but as a first approach we distinguish:

1. Lexical resources
2. Corpora with annotations per word form occurrence (token), such as lemma, pos, form, extended pos, (maybe also) chunk annotations, etc.
3. Corpora with a full syntactic structure for each sentence
4. Corpora with other types of annotations (to be investigated)

Class 1 includes CELEX, CGN lexicon(s), Cornetto, Open Source Dutch Wordnet, Duelme-LMF, GTB, GrNe, …

Class 2 includes inter alia CGN, SONAR-100, SONAR-500 and SONAR New Media, VU-DNC, Childes Corpora, FESLI and other SLI databases, VALID databases, Basilex, Nederlab data, MIMORE data, Corpus Hedendaags Nederlands, Corpus Gysseling, etc. ; SHEBANQ for Biblical Hebrew;

Class 3 include the CGN-treebank, LASSY-Small, LASSY-LARGE, SONAR Treebank, treebanks created with PaQu, and CHILDES treebanks in production by UU.

Class 4 include certain annotations in VU-DNC, Discan, possibly some annotations in CHILDES corpora, UU learner corpora, UU and other correctness corpora.

All text corpora (and textual transcriptions of pictures, audio- and audio-visual corpora) can in addition have a hierarchical grouping in terms of sentence/utterance, paragraph, section, etc. and these units may have metadata associated as well (e.g. the speaker of an utterance) etc.

Most resources contain linguistic annotations of one type (e.g. only morphosyntactic and syntactic annotations), but for several queries additional information from different linguistic levels is required (e.g. morphological, phonological or semantic information). For such cases a filtering of the search/analysis results on the basis of information on these linguistic properties that can be derived from others sources (in particular, lexical resources) is required. For concrete examples, see [Odijk 2014, slide 52]. We call this *chaining search*.

OpenSONAR already supports *batch queries*. What is actually desired is *parameterized batch queries*. In a parameterized batch query a normal query is made but it contains parameters. If a query contains *n* parameters, the parameters are instantiated by values from a list (hence: batch) of *n*-tuples of values. All local search engines should be extended with this functionality and federated search should support it as well.

All search applications should support the manual annotation of search results with codes to further refine the search results (cf. the Lancaster BNC interface).

[Kemps-Snijders 2015 Task 4.2] proposes a Multitier Annotation Search prototype with a single aggregated index and search engine and interface for all (textual) corpora and annotations types. Though it makes sense to investigate such an option if only for efficiency reasons, the approach has inherent risks, which should be taken into account: it is not obvious that one solution for all types of search will work or will yield user experiences that are comparable to dedicated search engines with dedicated user interfaces for one annotation type, such as GrETEL or OpenSONAR. Everything in one location has advantages but also disadvantages (e.g. restricted to the capacity that a particular centre can offer). The approach is based on KorAP, which itself is still under development, and which uses a query language (ISO CQLF) which is also still under development and whose expressivity is unclear (the ISO pages provide no information on ISO CQLF at all). For these reasons we should not gamble only on this one horse, and the federated approach with updates of dedicated search engines that already have proven their successfulness has to be pursued in parallel. Furthermore, it seems wise to phase this activity by starting with a proof-of-concept phase with a limited investment, and only to invest more after a positive evaluation.

[what we have]

- Local search
    - Class 1: Cornetto, DueLME-LMF, GTB, CELEX search engines, LEXUS, CGN?; Class 1 but of a different type: GrNe
    - Class 2: For most of these, dedicated search engines exist (COAVA for (part of CHILDES), FESLI for SLI data, TROVA for CGN, OpenSONAR for SONAR-500 and SONAR New Media, Trova for Childes Corpora, Nederlab Search (under development), MIMORE, Corpus Hedendaags Nederlands, Corpus Gysseling, SHEBANQ), but not for SONAR-100, VU-DNC, Basilex, VALID databases. Most can deal with the CQP language as query language. For Basilex and VU-DNC Auto-Search (first version under development at INL) might yield a search engine for these corpora, since both are in FOLIA format.
    - Class 3: GrETEL, GWR/PaQu are the most important search engines. All use XPATH/XQUERY as the query language.
    - Class 4: unclear (ANNIS?)
- Federated search: some end points but only supporting SRU/CQL

[what is newly needed]

- Local search: extend **[T47]** PaQu with combined data and metadata search and analysis, **[T40]** extend GrETEL with new interface and with analysis options of search results;
- **[T48]** An aggregated index search engine based on KorAP that allows users to upload new annotations and is integrated with NederLab.
- **[T49]** Extend Autosearch with further combined data and metadata search and analysis
- **[T35]** Federated search: see [Kemps-Snijders 2015] Tasks 4.1-4.3;
- **[T36]** Chaining search: e.g. chaining OpenSONAR/GrETEL with search in Cornetto (semantic information) and CELEX lexicons (phonological and morphological information)

- **[T37]** All search applications should be extended with facilities for the manual annotation of search results with codes to further refine the search results (cf. the Lancaster BNC interface)

[updates/upgrades]

- Local search: Most local engines require updates because actual usage has shown bugs and/or lacking important functionality. This holds for **[T39]** MIMORE [Barbiers 2014], **[T40]** GrETEL [Odijk 2014d] and **[T41]** OpenSONAR [Odijk 2015]. **[T76]** Engines and interfaces may have to be extended to support for local search for other properties
- **[T35]** Federated search: see [Kemps-Snijders 2015] Tasks 4.1, 4.3, 4.4; **[T36]** chaining search.
- **[T43]** CELEX web service and web application should be hosted by INL (instead of MPI/TLA). [Odijk 2014c Task 12]

[Parties] Meertens (generic), INL (generic,CELEX,) RUN (Basilex), UU (CHILDES treebanks, upgrade GrETEL), RUG (PaQu)

[References]

- **Local search:** [Odijk 2014c] tasks 1, 6,12, 13
- **Restricted Federated Search:** [Odijk 2014c] tasks 3,4,5,7; [Kemps-Snijders 2015]
- **Extended Federated Search:** [Odijk 2014c] tasks 8,9,10
- **Upload and Search:** [Odijk 2014c] tasks 2
- **Chaining Search:** [Odijk 2014c] task 11
- **CELEX** [Odijk 2014c] task 12
- **Taalportaal queries:** [Odijk 2014c] Task 19

2.4.3 Analysis

[should contain] Facilities to analyze data, in part inside search engines (to analyze the search results), in part existing external analysis tools such as R, SPSS, etc. Dedicated tools for linguistic, stylistic, readability analysis

[what we have] Some existing content search engines have some analysis functionality, but generally quite limited (PaQu, OpenSONAR), and some have hardly any options for analysis of the search results (GrETEL). Gabmap; Stylene (Flanders) for stylistic analysis, T-Scan (UU/Nijmegen, currently not part of the CLARIN infrastructure) for readability analysis (both not fully CLARIN/CLARIAH compliant yet)

[what is newly needed

[updates/upgrades] All analysis functionality of all search engines has to be upgraded, to incorporate combinations of multiple data and metadata elements, including own manual annotations of the search results. See [Odijk2015] for **[T41]** OpenSONAR, and desiderata for **[T47]** PaQu and **[T40]** GrETEL are known but still have to be put on paper.

[Parties] INL (AutoSearch, OpenSONAR), UvT (OpenSONAR), RUG (PaQU), UU (GrETEL), Meertens (generic analysis services)

[References] [Odijk 2015] [Odijk 2014c Task 15, 20] [Kemps-Snijders 2015 Task 4.4]

2.5 Visualisation

[should contain] Should offer visualization options integrated in the search / analysis tools or as separately callable functions. These options should be compatible with the general approach to visualization to be developed in WP2, for which the plan written by the Visualisation Expert Group coordinated by Erik Tjong Kim Sang forms an excellent basis [Tjong Kim Sang 2014].

[what we have] see [Tjong Kim Sang 2014].

[what is newly needed] Nothing specific for visualization will be done. Visualisation will be used inside applications.

[updates/upgrades]

[Parties]

[References] [Odijk2014c Task 15]

2.6 Data and Software 'Publication'

2.6.1 Data Publication

[should contain] Facilities and procedures to offer data to a CLARIAH centre, facilities to assign PIDs to the data and their metadata, to store the data and metadata on an accessible CLARIAH centre server, to publish the (CMDI) metadata via OAI-PMH and as LD (e.g. RDF triples via a SPARQL endpoint), to have the metadata harvested by metadata service providers such as VLO for browsing and searching, and whatever will be made in WP2 for CLARIAH. Facilities to upload the data in a search engine for local search and for federated search via this search engine's search end point. Facilities for proper arrangement of IPR and ethical issues. See [Brugman 2015 Task 1.5]. Facilities for long term storage (archiving) [Brugman 2015 Task 1.4]

[what we have] Some facilities and procedures to offer data to a CLARIAH centre, facilities to assign PIDs to the data and their metadata, to store the data and metadata on an accessible CLARIAH centre server, to publish the (CMDI) metadata via OAI-PMH, to have the metadata harvested by metadata browsers and search engines such as VLO, and whatever will be made in WP2 for CLARIAH. A conversion tool from CMDI to RDF has been developed in CLARIN-NL (in cooperation with CLARIN Austria). Facilities to upload the data in a search engine for local search and for federated search via this search engine's search end point. Facilities for proper arrangement of IPR and ethical issues. Meertens and INL are both recognized CLARIN centres and have the DSA, so long term storage is probably properly arranged but should be aligned with WP2 CLARIAH procedures and cost covering.

[what is newly needed] [T51] Making LD metadata (e.g. as RDF triples accessible via a SPARQL endpoint). Facilities to upload the data in a search engine for local search and for federated search via this search engine's search end point will be covered elsewhere (see section 2.4.1).

[updates/upgrades] [T52] Facilities for proper arrangement of IPR and ethical issues. **[T53]** Facilities for making archiving easier / more automated are desirable (cf. TLA's LAT and DANS's EASY).

[Parties] Meertens, INL, and via WP2 also other CLARIAH centres

[References] [Brugman 2015 Task 1.5]

2.6.2 Software deployment and hosting

[should contain] facilities offered by CLARIAH centres or other organizations (e.g. SURFsara) to host and deploy services. Policy for deciding whether to keep a service `live' or putting it to rest in an archive. Policy for versioning, preferably PIDs associated to services/web applications. Service virtualization options have to be formalized and well documented and also a proper administration interface with UI has to be provided. **[T79]** Development of a deployment framework (deployment procedure, administration of available services, etc.). This has to be done in close cooperation with WP2.

[what we have] **to be added**

[what is newly needed]

[updates/upgrades] **

[Parties] Meertens, INL, and other CLARIAH centres

[References]

2.7 Enhanced publications

[should contain] facilities to publish articles together with data and tools. This must be done in close collaboration with actual publishers (Lingua has claimed to be very interested) and with university libraries. Ideally we work with Open Access Journals, but having a close successful cooperation with a commercial publisher of a highly esteemed journal could serve as a leading example. Close collaboration with WP2 and all other WPs is required since everybody will have the similar needs and will run into the same problems. Experiments such as CLIO-DAP did but now in the linguistics domain and with linguistic journals are desirable.

[what we have] DANS experimented with this in the CLARIAH-SEED CLIO-DAP project, mainly in the area of socio-economic history. The workflows and procedures developed there should have wider applicability. We probably can learn from experiences in other disciplines as well.

[what is newly needed] No effort or budget is allocated for this.

[updates/upgrades] unknown

[Parties]

[References]

3. Cooperation with Other WPs

Cooperation with WP2 has been extensively mentioned above. Clear arrangements have been made between WP2 and WP3 on who is going to do what.

Cooperation with other WPs. In particularinvolves **[T58]** experiments to extract structured information from textual material by means of language technology in combination with the semantic interoperability facilities created in WP2 and the other WPs. This is potentially important both for socio-economic history (e.g. extract socio-economic information from news text) and for media studies (e.g. for automatically formalizing metadata information). It has been agreed with the other WPs to cooperate on this together, with investments from each. Topics being considered are in the domain of the Athena project and involve extraction of structured information of 'filmladders'. [Odijk 2014c Task 16,17], [Vossen 2015 Task 2.5]

4. References

[Barbiers 2014] S. Barbiers (2014), Desiderata MIMORE, Ms Meertens Institute

[van den Bosch 2015] Theme 3: Enrichment and Annotation

[Brugman 2015] Theme 1: Data and Metadata

[Kemps-Snijders 2015] Theme 4: Search and Research

[Odijk 2013] Odijk, J. (2013) 'The CLARIN Infrastructure (NL Part):Current State and Near Future', key note presentation held at the Digital Humanities Summer School, Leuven, September 20, 2013.

[Odijk 2014a] Odijk(2014), Gebruikersvriendelijkheid en Interoperabiliteit. CLARIN-NL memo met bijlage. Ms Utrecht

[Odijk 2014b] Odijk, J. (2014) Discovering Resources in CLARIN: Problems and Suggestions for Solutions, Ms. Utrecht University.

[Odijk 2014c] Odijk, J. (2014) CLARIAH Linguistics Projects Excel Work book version 2014-11-23.

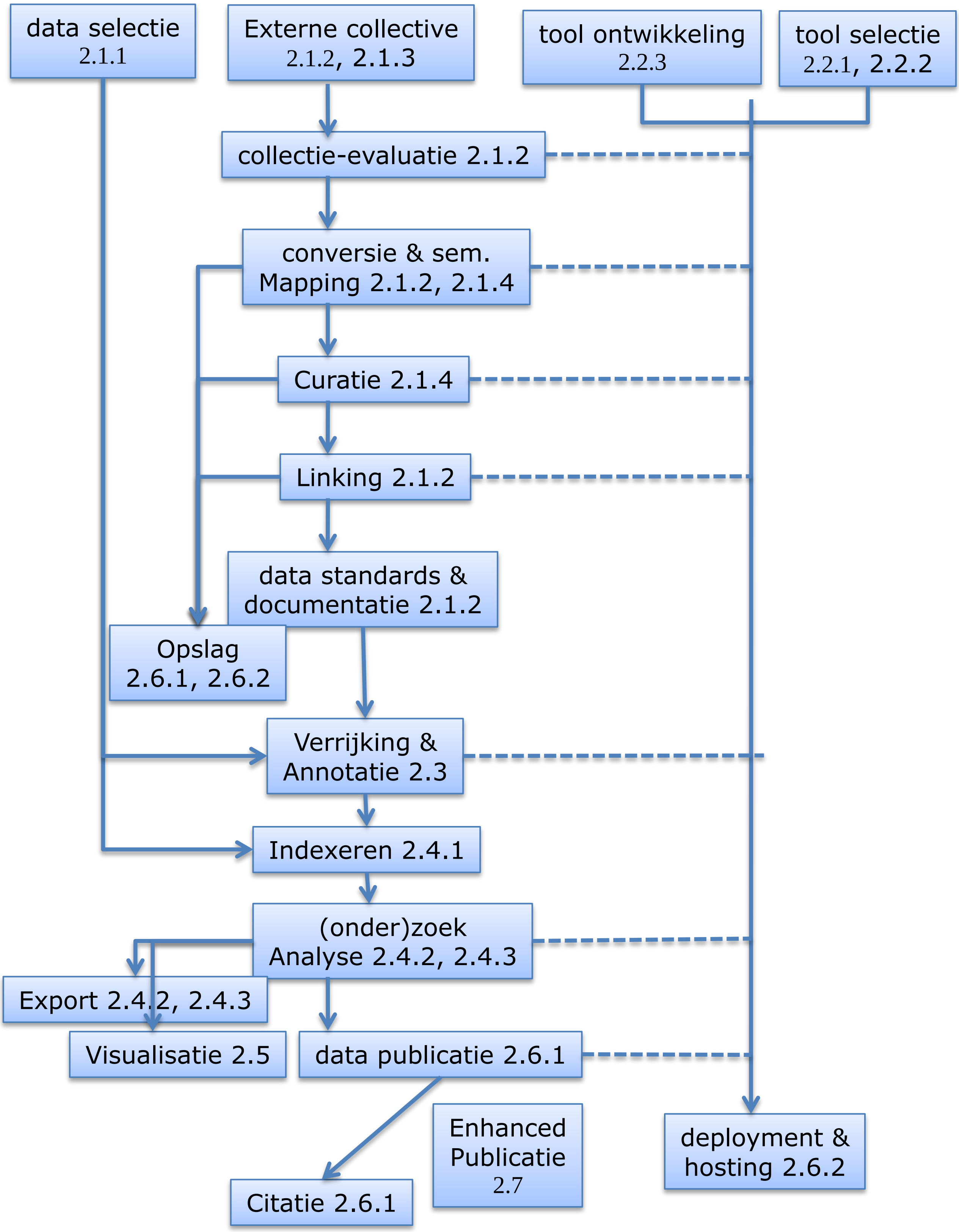
[Odijk 2014d] Odijk, J. (2014) Suggesties voor additionele functionaliteit in GrETEL. Ms Utrecht University

[Odijk 2015] Odijk, J. (2015), Wensenlijst OpenSONAR, CLARIN-NL internal document, 2015-02-04.

[Tjong Kim Sang 2014] Tjong Kim Sang, E. (2014) *CLARIAH: Plan: Analysis and Visualisation*

[Vossen 2015] Theme 1: Interoperability

5. Appendix CLARIAH onderzoeks-workflow


data selectie
2.1.1
Externe collective
2.1.2, 2.1.3
tool ontwikkeling
2.2.3
tool selectie
2.2.1, 2.2.2
collectie-evaluatie 2.1.2
conversie & sem.
Mapping 2.1.2, 2.1.4
Curatie 2.1.4
Linking 2.1.2
data standards &
documentatie 2.1.2
Opslag
2.6.1, 2.6.2
Verrijking &
Annotatie 2.3
Indexeren 2.4.1
(onder)zoek
Analyse 2.4.2, 2.4.3
Export 2.4.2, 2.4.3
Visualisatie 2.5
data publicatie 2.6.1
Enhanced
Publicatie
2.7
deployment &
hosting 2.6.2
Citatie 2.6.1