Originally published in: Fišer, Darja/Witt, Andreas (Eds.): CLARIN. The infrastructure for language resources.
Berlin/Boston: de Gruyter, 2022. Pp. 307-339. (Digital Linguistics 1)
DOI: *https://doi.org/10.1515/9783110767377*

Piotr Bański* and Hanna Hedeland

Standards in CLARIN

Abstract: This chapter looks at a fragment of the ongoing work of the CLARIN Standards Committee (CSC) on producing a shared set of recommendations on standards, formats, and related best practices supported by the CLARIN infrastructure and its participating centres. What might at first glance seem to be a straightforward goal has over the years proven to be rather complex, reflecting the robustness and heterogeneity of the emerging distributed digital research infrastructure and the various disciplines and research traditions of the language-based humanities that it serves and represents, and therefore part of the chapter reviews the various initiatives and proposals that strove to produce helpful standards-related guidance. The focus turns next to a subtask initiated in late 2019, its scope narrowed to one of the core activities and responsibilities of CLARIN backbone centres, namely the provision of data deposition services. Centres are obligated to publish their recommendations concerning the repertoire of data formats that are best suited for their research profiles. We look at how this requirement has been met by the particular centres and suggest that having centres maintain their information in the Standards Information System (SIS) is the way to improve on the current state of affairs.

Keywords: standards, formats, CSC, SIS, data deposition

1 Introduction

This chapter looks at the ongoing work of the CLARIN Standards Committee (CSC) on producing a shared set of recommendations on standards, formats, and related best practices supported by the CLARIN infrastructure and its participating centres.

Acknowledgment: We are grateful to the members of the CLARIN Standards Committee for their participation in the process that resulted in publishing the re-vamped Standards Information System, and for their support and sharing ideas on how it can be made better. Very special thanks are due to Eliza Margaretha Illig, the main coder of the SIS, for her enthusiastic and professional participation in the project. Hanna Hedeland's work was funded by the BMBF-project QUEST (16QK09D) at the Leibniz-Institut für Deutsche Sprache.

*__Corresponding author: Piotr Bański,__ Leibniz-Institut für Deutsche Sprache, Mannheim, Germany, e-mail: banski(at)ids-mannheim.de
Hanna Hedeland, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, Germany, e-mail: hedeland@bbaw.de

https://doi.org/10.1515/9783110767377-012

What might at first glance seem to be a straightforward task has over the years proven to be rather complex, reflecting the robustness and heterogeneity of the emerging distributed digital research infrastructure and the various disciplines and research traditions of the language-based humanities that it serves and represents.

In late 2019, the CSC decided to reduce the initial scope of the task in order to make it both manageable and immediately relevant to the current needs of the CLARIN community. The focus was therefore narrowed to one of the core activities of CLARIN centres, namely data deposition services, and stress was placed on a measurable requirement concerning the so-called B-centres, namely the publication of each centre's recommendations concerning the repertoire of data formats that are best suited for deposition at that particular centre. While it is more restricted than the original goal, and thus more tangible, this smaller task requires a careful balance between the top-down across-the-board demands of a modern distributed research infrastructure, and the bottom-up expression of the research orientation of the particular nodes in the network, that is, the individual centres. It also requires the formation of an inventory of formats and ways of evaluating them for appropriateness as shared recommendations. Another goal that must be met in order to address the task is that of ensuring sustainability and ease of maintenance of the proposed solutions, while at the same time ensuring that these solutions will become a useful tool – for the CLARIN staff, both in the centre-assessment process and as a source of developer-oriented detailed information on data formats, and also for the users of CLARIN who wish to deposit data, to assist them in the task of identifying centres that best suit their needs. The emerging system, in the next step, will serve the larger goal of gathering information on the major relevant standards used across CLARIN, as well as other related research infrastructures.

In the remainder of this section, we first define the scope of the present chapter, and then outline its structure.

1.1 Scope

For the purpose of this chapter, we differentiate between

(a) standards, which are the result of a formalized standardization process and are published by a standardization body, such as ISO, W3C, OASIS or others;
(b) (data) formats, which may be a serialization of a standard, but where the only requirement is a reliable specification or schema;[1] and

1 For an interesting discussion of several possible definitions of the somewhat narrower term 'file format', in the context of sustainability assessments, see, among others, (Pennock, Wheatley, and May, 2014).

(c) best (or good) practices, which are formats and *de facto* standards that are generally accepted as the (or *a*) recommended solution for a particular method or context, considering both the usage of, and the tool support for, the given format, as well as its features.

Given the complex nature of the task of defining a shared set of recommendations, the space and time restrictions of this publication, and the fact that the work of the CSC is far from finished, we narrow the focus of the present chapter to data formats. Our special attention here is on the implementation of a flexible and maintainable solution based on reliable transparent workflows for revisions and quality control to ensure that CLARIN is able to respond appropriately to relevant future development within and beyond the infrastructure.

Furthermore, we focus entirely on CLARIN, to the exclusion of other projects or research infrastructures. Due to our own backgrounds and the composition and activities of the CLARIN Standards Committee, our perspective will inevitable also be somewhat tied to the German consortium, CLARIN-D.

1.2 Structure

In what follows, we first sketch the theoretical and institutional background for the activity of the CSC (Section 2), and after that, we look at the history of the struggle to flesh out standards-related guidelines for CLARIN researchers and users (Section 3). In Section 4, we present the formal factors that influence the task at hand, and in Section 5, we show how the CSC has addressed it, culminating in the re-emerging Standards Information System. We finish with a summary and indication of directions for the next steps.

2 Background

Within CLARIN's designated communities, there exist, on the one hand, users whose work results in the development of new standards and formats that are later adopted by others, and, on the other hand, users who are unable to make a suitable choice from existing standards and formats for their own research project. The extreme variation in data literacy often accompanies methodological differences, and additional dimensions are introduced due to different linguistic modalities, research areas, and traditions. This heterogeneity results in a plethora of standards, formats, and localized best practices in use within CLARIN and asso-

ciated institutions. In order to handle such a situation, an infrastructure would need highly specific expertise in an increasing number of areas. Fortunately, each centre joining the CLARIN project contributes substantial and often innovative expertise based on their own research and the needs of their designated communities. While this is undoubtedly one of the strengths of a distributed infrastructure, it also implies that centres may have differing views on both their own and their users' needs when it comes to shared recommendations and other kinds of support in matters concerning standards and formats. And that calls for solutions that respect and embrace heterogeneity without confusing it with arbitrariness. Some established formats exist alongside very similar – possibly more modern – formats, due to minor yet crucial differences in expressiveness, superior tool support or local habits, and so on, and in many cases this can only be recognized with highly specific expertise within the relevant area. Therefore, any centralized or otherwise non-representative decision-making process in producing a set of shared recommendations on standards and formats will inevitably fail to receive the necessary support from the partners – and users – of CLARIN.

In this section, we first look at how CLARIN deals with the heterogeneity that is implied by its structure (Section 2.1), and then, in Section 2.2, we present requirements concerning data and services that are generally accepted across modern research infrastructures and that act as a top-down framework that prevents heterogeneity from becoming chaos.

2.1 Heterogeneity and interoperability

While CLARIN as a whole benefits from the expertise of individual centres, the converse is also true: interconnected CLARIN centres also benefit from being part of the infrastructure, both as institutions and with regard to what they can offer to their users. Certified CLARIN B-centres accepting digital resources can ensure long-term archiving by their own means, but can also be supported by other CLARIN centres, should one centre run into funding problems or even be forced to shut down completely. The common infrastructure also includes services that a single centre could never provide, such as the Virtual Language Observatory (VLO)[2] (Windhouwer and Goosen, 2022), the Federated Content Search (FCS)[3] (Schonefeld et al., 2014; Olsson, 2017), and the Language Resource Switchboard[4]

2 https://vlo.clarin.eu/
3 https://contentsearch.clarin.eu/
4 https://switchboard.clarin.eu/

(Zinn and Dima, 2022). Services on a national level, for example, the German WebLicht web service orchestration platform (Hinrichs, Hinrichs, and Zastrow, 2010), the LINDAT/CLARIAH-CZ web services (Hajič et al., 2022) or the PORTULAN Workbench (Gomes et al., 2022), can also aggregate the efforts of several centres or institutions and become discoverable beyond the national context through the CLARIN infrastructure.

In these cases, CLARIN has shaped its own best practices: for example, centres are required to provide metadata in the CMDI format (Broeder et al.,2012; Goosen et al., 2015; Windhouwer and Goosen, 2022) for the resource portal VLO, and although the FCS uses generic standards such as the query protocol Search/Retrieve via URL (SRU)[5] and the Contextual Query Language (CQL)[6] to enable searching in collections across the infrastructure, centres also comply with additional CLARIN FCS specifications for querying language resources on various levels. The development of such common CLARIN-specific practices and procedures has been achieved by the respective task forces of the Standing Committee for CLARIN Technical Centres (SCCTC). In contrast with services like the VLO and the FCS, for which centres provide resources and users interact with controlled GUIs, the situation is much more challenging when users are allowed to interact directly with services such as WebLicht or the Language Resource Switchboard using their own data, which comes in various formats, or when users are generally looking for tools and services to implement their data creation and analysis workflows.

CLARIN cannot and should not support all conceivable formats, but rather a well-defined subset[7] including *de facto* standards and formats relevant to the respective disciplinary and data communities (cf. Cooper and Springer, 2019). One of the first steps in specifying any measure of CLARIN-wide guidelines is therefore not only to review what these formats are, but also to make clear why certain formats should not be supported by CLARIN, even though some centres might still need to accept them.[8] This is a way of gently pointing users towards formats that comply with the current data quality requirements (see Section 2.2), thereby avoiding immense data curation costs.

5 https://www.loc.gov/standards/sru/

6 https://www.loc.gov/standards/sru/cql/

7 A more comprehensive, fine-grained approach to categorizing format impact, defining several levels of interoperability based on the status of formats ranging from internationally recognized or *de facto* standards and best practices, via formats and standards that are only regionally relevant or discipline-specific, to less prioritized and more rarely used formats, is outlined in Odijk (2016).

8 Depending on the research profile and target data, this is indeed the case in some centres, where the value of the donated data outweighs up-translation costs, cf. Thomas and Wiegand (2015).

To arrive at transparent, widely acknowledged recommendations reflecting the situation in individual CLARIN centres and their designated communities, a crucial requirement is to understand the functions and roles of the various formats in the research process and as parts of complex resources. When assessing individual formats, it is just as crucial to differentiate between, on the one hand, aspects that reflect research traditions or theories, which become visible through data modelling decisions, and, on the other hand, aspects that are not defined by, or relevant for, the research process, but nevertheless vary across formats. An example of the latter would be various ways of modelling alignment between a recording and a transcript that are not directly relevant from the perspective of researchers using their customary tools and formats, as opposed to different options for annotation structure and schemes that directly affect the way in which research questions and analyses can be expressed (cf. Schmidt, 2011). Such a task is by no means trivial. The expertise and experience accumulated within CLARIN offers a unique opportunity to arrive at the appropriate solutions and to provide researchers with the information they need to create better data. Accumulating detailed and qualified information on recommended and used formats, and appropriately exposing and visualizing that information for the purposes of querying and comparison, makes it possible to move forward and enhance interoperability across the infrastructure.

2.2 Quality criteria for data formats

When it comes to research data, quality criteria go beyond assessing generic format sustainability, although the latter is always required as a baseline. For this generic type of sustainability assessment, several organisations provide information, guidelines, and metrics (see Section 5.2 for examples). Even if, until recently, the criteria for basic research data quality and best practices were not entirely clear, today the FAIR principles, which require data to be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016), have become common ground among initiatives related to research data management. At the same time, the idea of machine-actionable data with a well-defined semantic model promoted with these principles is new to most of the humanities, and maybe even out of reach according to some (RDA FAIR Data Maturity Model Working Group, 2020, 10).[9]

9 "[D]ata coming from humanities fields, especially from outside of Digital humanities, will often not be expressed in a machine understandable knowledge representation (RDF, SKOS or LOD) by nature but instead, it is often expressed in natural language, even if encoded using machine readable methods (e.g., TEI). Therefore, it becomes quite clear that the indicator treating machine-

Many formats traditionally used in the Humanities, for example, formatted text documents, are indeed not even reliably machine-readable or processable. Due to their nature as domain-independent, high-level principles, the FAIR principles do not offer direct guidance on actual formats. The idea is that they serve as the basis of an implementation process for a specific discipline and/or context.

Implementation of the FAIR principles within the CLARIN infrastructure has come a long way (de Jong et al., 2018, 2020), and the technical and administrative means are in place to guarantee that resources in certified CLARIN B-centres are findable and accessible. Thanks to advanced solutions for metadata, PIDs (Persistent Identifiers), and AAI (Authentication and Authorization Infrastructure), these first two aspects of FAIR, which are not directly related to the resources themselves, are already fulfilled. However, when it comes to the requirements that the data should be interoperable and reusable, the technical infrastructure used for the safeguarding and distribution of research data is not in a position to fulfill these, as they to a large extent depend on characteristics of the deposited data itself. While many CLARIN resources are undoubtedly among the FAIRest of their kind, there is still work to be done to ensure interoperability that goes beyond format conversion and syntax. In order to enhance data interoperability and reusability, resources need to be understandable to both humans and machines. Therefore, the semantics of data formats and the schemes and conventions used within these formats have to be taken into consideration. Established "domain-relevant community standards" (Principle R1.3, Wilkinson et al., 2016) for data and metadata are still lacking, for example, in the area of (Linguistic) Linked (Open) Data (cf. Chiarcos, Fäth, and Abromeit, 2020). The technical and methodological expertise of CLARIN together with the expertise and needs of its users from various research and data communities will allow for a successful evaluation and further development of relevant data formats based on the FAIR principles.

3 Evolving standards recommendations in CLARIN

This section briefly outlines the context and results of previous initiatives that led to the approach described in the present chapter. One has to bear in mind that the concept of a distributed digital research infrastructure for the language-based humanities is novel in its nature, and both technical and governance solutions

understandable knowledge representation will be less relevant according to the Humanities." (RDA FAIR Data Maturity Model Working Group, 2020, 10)

have been emerging over the years and are still being developed in a natural process of maturation. Still, the need for a common set of recommendations concerning standards to be used in CLARIN was already obvious in the preparatory phase of the project. This resulted in several takes on the issue and eventually several sets of recommendations, differing in their structure, granularity, outreach and authority, although all of them seem to have made a single assumption about such a list: namely that it can be established centrally and that it can be effectively imposed on the centres and users in a top-down fashion.

At the beginning of the CLARIN project, research data deposits were not common. With the increasing digitalization and datafication of society in general, awareness of topics related to research data management, and funders' requirements to deposit data for scientific reuse whenever possible, a cultural change was initiated, and is still very much in progress. And with the increased amount of data available, the focus has turned from building technical solutions for the "F" and "A" in FAIR to the data itself, that is, to the "I" and the "R", and questions of data quality (RfII, 2020). As certain centres experienced an increase in deposits, it became clear that these experiences must be continuously integrated into the corresponding recommendations – and that, conversely, these recommendations must be available to users who are interested in creating or depositing resources complying with current good practice.

In 2015, someone looking for CLARIN recommendations on standards and formats would face a number of partly contradictory sources (an extensive list of the documents and other sources that punctuated the project timeline can be found in Annex A to this chapter). In the German use case, the CLARIN-D Data Management Wizard was based on incomplete sources and therefore also omitted formats accepted by several German centres. The "User manual for CLARIN-D" referenced from the wizard referenced in turn the "Standards for LRT" document from 2009 with different information. At the same time, on the CLARIN website, there was information on some centres' format preferences on the "Standards and formats" page, but also another list of standards in a FAQ titled "What standards are recommended by CLARIN?". Around 2012, the CLARIN-D centre at the IDS also provided the CLARIN Standards Guidance (a predecessor of the SIS, to which we turn in Section 5), which was technically advanced, interactive, and user-friendly, but based on incomplete and by then already somewhat outdated information. In 2013, the German funder DFG published a set of recommendations on technical aspects of the creation of language corpora, which was partly based on the (work leading to the) document CLARIN-D5C-3. The latter was not publicly available, and that also applies to the English translation of the DFG recommendations created later in 2015. There were also internal sources; in particular, the CLARIN document "Relevant data formats" (Van Uytvanck, 2014)

describes exactly what steps needed to be taken in order to arrive at a list of relevant formats for CLARIN and even references a spreadsheet (Annex A, 8.) with an initial list of formats and columns for information on their purpose and CLARIN-wide (top-down) level of recommendation. The document also makes clear the benefits that such an aggregated list would bring to several areas of the infrastructure, but the initiative was sadly not followed up. At their meeting at the annual conference in Wrocław, in late 2015, the Standards Committee decided to undertake the task of producing an up-to-date list in a bottom-up manner.

In our stall at the Bazaar of the CLARIN Annual Conference 2018 (Hedeland and Bański, 2018), an attempt was made to gather information directly from centre staff about their current practices and general preferences in recommending or discouraging formats. The discussions showed that the task was not only a matter of logistics and endurance in eliciting the information. Some centres rejected the idea of recommendations altogether, arguing that format-related preferences were something to be decided not by an infrastructure, but by individual researchers in accordance with their needs, and that attempts to regularize them might limit freedom of research. Recall that CLARIN serves very different users, highly skilled computer linguists as well as non-tech conversation analysts who are simply trying to comply with the funders' newest requirements. In both cases, standardization can only be implemented by abstracting away the theory-ladenness of research data formats and only applying recommendations to those aspects that are not affected. It also should be stressed that a list of supported formats will never imply that individual CLARIN centres should not accept additional FAIR-compliant data formats, or legacy data in discouraged formats. Guidance is, however, necessary for researchers who need support in creating high quality FAIR research data and to enhance interoperability within the CLARIN infrastructure.

4 Format recommendations: Assessment conditions and metrics

The backbone of CLARIN is composed of B-centres (service-providing centres; see Wittenburg et al. (2020) for more details), one of the primary roles of which is ensuring the longevity and curation of data that users may deposit with them. This section looks at how the relevant obligations of B-centres are specified in the certification requirements, fulfilled by centres, and used as one of the performance metrics of CLARIN.

4.1 Format-related assessment requirements

The assessment and certification of CLARIN centres is handled by the CLARIN Assessment Committee (CAC), and part of that process requires that centres are certified with the CoreTrustSeal (https://www.coretrustseal.org/, CTS for short).

As Wittenburg et al. (2020, 3) state in their CLARIN centre description, "Centres need to have a proper and clearly specified repository system and participate in a quality assessment procedure as proposed by the CoreTrustSeal." Wittenburg et al. (2019, 1), in a checklist document for centres that are candidates for type B, strengthen this requirement by stating that "[t]he centre cannot be certified as a B-centre until the CoreTrustSeal assessment has been successfully concluded (. . .) The application for the CoreTrustSeal, or proof that the CoreTrustSeal has been awarded, has to be provided."

The CTS requirements concerning formats, listed in the "Extended Guidance" document (CoreTrustSeal Standards and Certification Board, 2019), Section 8: "Requirements/Appraisal", are as follows:

> For this Requirement, responses should include evidence related to the following questions:
> (. . .)
> – Does the repository publish a list of preferred formats?
> – Are checks in place to ensure that data producers adhere to the preferred formats?
> – What is the approach towards data that are deposited in non-preferred formats?
> (. . .)

Of these questions, it is the first one that we focus on in the present chapter. In the remainder of this section, we look at how centres have addressed the requirement to publish lists of preferred formats, show how the degree of fulfillment was measured, and list features desirable in a system designed to assist in aggregating and visualizing the relevant information, while minimizing the effort needed to keep it current.

4.2 Addressing format-related assessment requirements

The CTS and thus B-centre-assessment requirements reviewed above provide a reasonably clear and measurable framework for centres to fit in. A KPI (Key Performance Indicator) has been established that measures the "percentage of centres offering repository services that have published an overview of formats that can be processed in their repository" (Maegaard and Wessels, 2019).

In theory, due to the assessment process, this KPI should be close to 100%, potentially deviating from the maximum only in the case of non-B-centres that allow for data deposition but are not regulated by the CTS, or, marginally, in the case of B-centres that are in the process of reassessment.

In practice, centres have adopted various strategies to address the CTS requirements: some centres have indeed published lists of recommended formats,[10] with their own subdivisions and varying granularity, and various ways to indicate their interest in receiving various formats, while other centres, possibly as an expression of their readiness to accept any data in nearly any format, have directed users towards the previously announced CLARIN top-down recommendations, most notably the "LRT standards" document.[11]

Another factor, pointed out to us in personal communication by Dieter Van Uytvanck, is that the above-mentioned requirement for B-centres to provide deposition services has acquired a fuzzy interpretation that sometimes invokes "internal deposition" as a way to satisfy the assessment procedure. We do not take a stance here on the formal status of such an approach, merely noting it as another factor that influences centres' willingness to publish information about recommended formats.

In 2019, the CSC decided to focus on data-deposition format recommendations as a first step towards developing a list of standards recommendations in CLARIN that would be more modern and easier to maintain than the existing standards recommendations (see Annex A for a hopefully complete list). That decision led to the welcome consequence that the KPI rose from 33% reported in 2018 and 2019 to 46% in the following year.[12]

It has to be borne in mind that, for some colleagues responsible for addressing the CTS requirements, the issue is tied to freedom of research or the need to collect rare and valuable data at all costs. We believe that such an attitude is a natural consequence of the quasi-Platonic assumption that there exists a central

10 These centres can be found listed at https://www.clarin.eu/content/standards-and-formats.

11 These centres can be found listed at https://github.com/clarin-eric/standards/issues/14. A strong impetus towards recommending the "LRT standards" document came as a result of the precious initiative by LINDAT colleagues that unifies the information for data depositors and provides a FAQ that directs the reader to the "LRT standards" document. Current work on the SIS promises a replacement of that link with a centre-specific link to the recommendations (see Section 5.3 for an example).

12 The measurement reported to us is probably not perfect, because it does not take into account newly certified centres; what is important, however, is a significant rise in the percentage of centres publishing their own format recommendations; we are told by Dieter Van Uytvanck (personal communication) that a new round of KPI measurements is in progress at the time of writing.

top-down format recommendations list (even if the list is yet to be codified), and that format recommendations have an absolute, binary nature, not allowing for any form of gradation. We also believe that such an assumption should be eliminated and replaced with a more satisfactory system. The features of such a system are enumerated below:

1. There should be a way for a centre to publish format recommendations suited to and reflecting its own research profile, in such a way
   a. that the decision and the act of publishing can take place relatively quickly and painlessly,
   b. that the resulting collection can be updated quickly and straightforwardly.
2. These recommendations should ideally be structurally uniform across the board, to form a reliable basis that would enable users to select a centre for data deposition.
3. A new centre should be able to use a template, rather than devise yet another list.
4. Centres should not be tempted to "just link" to a single set of top-down recommendations, because those will rarely match a research profile (and are never meant to match a single profile).
5. The format taxonomy should be comparable (preferably, shared), and should ideally be able to also provide additional information (about comparable formats as well as about the standards documents that define many formats).
6. The results should be visualized in a way that allows one to glean extra information from the aggregation of the recommendations (e.g., about the most and least popular formats).

The following section shows that the upgraded Standards Information System meets the above description.

5 Standards Information System: Goals and description

The solution described in this chapter has arisen out of several sources: the general tension in the CLARIN community reflected in the Wrocław declaration of 2015, the stalled CLARIN Standards Guidance project and, more recently, discussions within the Standards Committee and the relevant part of the CLARIN KPI-related research.

Out of the above-mentioned factors, two have already been at least briefly touched upon in the preceding sections: the community tension (Section 3) and

the KPI-related research (Section 4). CLARIN Standards Guidance (Stührenberg, Werthmann, and Witt, 2012) was an early project meant to consolidate the repertoire of standards advocated by CLARIN (in a top-down fashion, by marking some of them as "recommended by CLARIN"). Despite being well-designed, based on modern XML and Semantic Web technology, and featuring useful visualizations, it became stalled due to the amount of work that its maintenance by a small team would involve, and effectively made the list of "previous standards collections" that is the subject of Annex A, with an outdated fragment of it quoted until recently at one of the clarin.eu pages as yet another set of recommendations.

The KPI-related research within CLARIN has been described by Maegaard and Krauwer (2018) and Maegaard and Wessels (2019). A part of that research relevant to the beginnings of the present-day SIS concerns the indicator "Collection of standards and mappings" (Maegaard and Krauwer, 2018), with the accompanying measure defined as "Percentage of centres offering repository services that have published an overview of formats that can be processed in their repository", and gave the Standards Committee an opportunity to focus more narrowly on an issue that promised to be both practical and useful, and to constitute a seed for further work on the far-reaching goal of the CSC.

The last of the factors that contributed to the rebirth of the Standards Guidance as the Standards Information System is work of the CSC after 2015, punctuated by Bański (2018), a white paper circulated among the members of the CSC and other interested colleagues that contained ideas that were further polished into the current proposal, among them a crude function-based division of formats and a version of levels of recommendation, encoded as a parameter matrix.

The present section looks at the CSC research concerning the relevant KPI, then moves on to outline the concept and content of functional domains and levels of recommendation, finally focusing on the current SIS and on how it addresses the various needs outlined in Section 4.2 and elsewhere.

5.1 Data collection

The data that formed the initial core of the work of the CSC after mid-2019 was collected by Dieter Van Uytvanck in a spreadsheet designed to measure the format-related KPI and at the same time to check how popular certain formats were among the CLARIN centres. The spreadsheet consisted initially of format names (of varying granularity) and collected data from the initial seven centres that published their requirements concerning deposition formats (Bański, Hedeland, and Van Uytvanck, 2019). In the course of 2019 and 2020, the spreadsheet was extended thanks to the efforts of the CSC members, finally embracing all those

centres (whether of the B-status or not) that offered deposition services, expanding the number of format names and gathering them into format families (in many ways preceding the functional domains that are the topic of Section 5.2.1). Popularity of the particular formats was measured by indicating a "1" in cells where the format name row and the centre name column met, and then by calculating the number of occurrences of "1", with results relativized to a particular format family, so that text annotation formats would not compete with audio encoding formats. The results of these stages of the CSC work can be found in the early, internal, releases at https://www.clarin.eu/content/standards; a glimpse is also provided in CLARIN Standards Committee (2020).

While the initial work on the KPI spreadsheet was fruitful and moved the KPI to another level within a year, with the members of the CSC ensuring that many B-centres in their spheres of influence published their recommendations, it also became clear that the system of counting only "1"s for an occurrence of a format name in a format list was far from satisfactory, as it did not take into consideration domains of application of the given format, or the level of support that the given centre assigned to it. This inadequacy was eventually addressed by formulating a list of functional domains (Section 5.2.1) and encoding three levels of recommendation (Section 5.2.2), and in a longer perspective, by abandoning the KPI spreadsheet as insufficiently expressive and focusing on the Standards Information System as the locus for information on format recommendations as well as the tool for gathering that information from centres as a way to enable them to satisfy the assessment requirements in a comparable and sustainable way (Section 5.3).

5.2 Design of format recommendations

The transition to the relaunched Standards Information System required the definition of both a data model for more elaborate format descriptions than the ones in the KPI spreadsheet and a schema for adequately modelled format recommendations. In order to be accepted as a reasonable alternative to existing practices, format recommendations in the Standards Information System have to be at least as expressive as those currently provided by centres. On the other hand, to encourage the contribution of information, brevity and simplicity are crucial, especially regarding descriptions of additional formats. The CSC decided to focus on those formats that CLARIN is particularly suited to provide the relevant information about. This way, it is possible to avoid information gathering and management in parallel with existing generic initiatives such as the Sustain-

ability of Digital Formats Website[13] of the Library of Congress, the U.S. National Archives and Records Administration (NARA) Digital Preservation Framework,[14] or PRONOM[15] of the U.K. National Archives. These initiatives already provide comprehensive and detailed information on most widely used formats, including assessments of their sustainability – and they also use persistent format identifiers that can be referenced from less detailed descriptions in the Standards Information System. Apart from the more generic orientation of the format registries and assessments of these examples in comparison with the intended purpose of the Standards Information System, the former focus mainly on long-term preservation and sustainability of the formats, while for CLARIN, the aspect of interoperability within the technical infrastructure in its current state is also important. Format recommendations provided by centres also differ from format assessments in the sense that centres do not need to provide any explanations for their recommendations and preferences. For these reasons, it has been decided initially to restrict the data models of the Standards Information System compared to the detailed format descriptions available elsewhere, and to only incorporate the information currently required for the task at hand.

5.2.1 Functional domains

The CLARIN centres that published white lists of formats or format recommendations most often used content-oriented categories for the sake of structural clarity and in order to provide guidance for users. That was not fully adequate for two reasons: firstly, no uniform categorization was adopted across centres, and, secondly, it was often assumed that categorization was a secondary projection of the nature of the particular formats and thus merely grouped them into "families" of a sort. This is also the approach in the Summary Guide to Preferred Formats[16] by The Dutch Digital Heritage Network (NDE),[17] based on the PRONOM and NARA Digital Preservation Framework information, where archives and data centres in the Netherlands list their preferred formats. If format sustainability is the only requirement, that is a sensible approach, but when the range of *functions* of data used by CLARIN centres is taken into consideration, it becomes clear

13 https://www.loc.gov/preservation/digital/formats/
14 https://www.archives.gov/preservation/electronic-records/digital-preservation-risk
15 https://www.nationalarchives.gov.uk/PRONOM/
16 https://www.wegwijzervoorkeursformaten.nl/index.php/Summary_Guide_to_Prefered_Formats
17 https://netwerkdigitaalerfgoed.nl/

that the relationship between formats and functional domains must be many-to-many, because very often a single format can be (and is) used for more than one purpose. One common example is PDF/A, which is a highly recommended format for long-term archiving of, for instance, unstructured resource documentation such as annotation guidelines or a corpus manual, or for digitized scans of original texts. It is, however, seldom a recommended format for the resource itself, for example, for text annotations or audiovisual annotations. No common set of categories describing these functions has been in use in CLARIN, although a Resource and Technology Taxonomy including this type of information was drafted already in the preparation phase (Wittenburg et al., 2008); the current status of this draft or its impact on CLARIN centres remain unclear.

While the CLARIN Standards Committee was extending the KPI spreadsheet, the initial workaround was to focus solely on a single purpose: the use of formats for linguistic research data in the narrowest sense, such as text and annotations, while ignoring other format recommendations. The wish to reflect the greater and more complex picture, however, soon led to the development of a set of functional domains reflecting the relevant data types in CLARIN repositories. The initial set was based on the results of a survey of several repositories holding various types of resources, which was carried out in the project QUEST[18] at the IDS, complemented with the expertise and experience of the members of the CLARIN Standards Committee.

The proposed functional domains overlap with the draft taxonomy of Wittenburg et al. (2008) to a large extent, as would be expected from two descriptions of the same area. However, The Resource and Technology Taxonomy also includes some abstract categories such as Object, Situation and Session, which are not relevant to the format-oriented CLARIN Standards Information System. The taxonomy differentiates between speech (audio) and multimodal (video) resources, both of which belong to the functional domain Audiovisual Source Data, but it does not differentiate between annotations referring to audio or video resources on the one hand, and those referring to text on the other; in the SIS, these annotations are considered to represent different functional domains. Furthermore, in contrast to the taxonomy, the functional domains proposed here consider transcripts to be a subtype of annotation.

The currently identified set of functional domains is listed in Annex B. In the remainder of this section, some less obvious choices and categories are briefly explained. A fundamental assumption that has to be borne in mind is that the purpose behind using functional domains is not so much for them to constitute

18 https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html

a complete knowledge resource or ontology concerning data types or functions in the language-oriented humanities, but rather to allow the Standards Committee to elicit relevant information about standards and formats in use within the CLARIN infrastructure. The original set of domains is thus a first version that might be refined and extended according to additional requirements established through the actual use of the SIS. For each functional domain, there will most likely always exist a set of recommended formats – there are valid reasons why a single uniform exchange or standard format has not yet replaced all others – especially given that several formats already have a strong geographic or discipline-based support.

In the area of (structured) resource documentation, a distinction is drawn between the three categories: "Contextual Information", "Catalogue Metadata", and "Metadata". This distinction is not used explicitly in all centres, and the aim in providing three categories is to find out more about which highly related formats are used for which exact purposes. The reason for singling out information on texts or communicative events and authors or participants as "Contextual Information" is that this information (a) is highly dependent on the research question at hand, and can therefore never be standardized with regard to the elements and values used, and (b) can contain too much potentially sensitive information to be in the public domain. In CLARIN, the standard metadata format is CMDI (cf. Section 2.1), but one of the main aims of CMDI is the harmonization and standardization of metadata within a centre (and partly within CLARIN), and another is the public availability of the metadata records for harvesting. It is therefore expected that centres use, or at least handle, additional formats for richer and potentially non-public information. Furthermore, in addition to CMDI, which is required within CLARIN, many centres also provide reduced sets of metadata for resource discoverability in contexts other than the VLO, with Dublin Core[19] being the typical example of "Catalogue Metadata". This metadata only contains very basic discoverability information required for being listed in generic catalogues or portals for archives or research data centres of various types.

Other categories in the set are very broad, for instance, the category "Tool support", including all kinds of formats related to tools and services. While there are undoubtedly conceptual differences between a tagset, a language model, and a settings file including tier formatting information, for the purpose of gathering information on formats, further subcategories seem unnecessary, at least in the

19 https://www.dublincore.org/

initial stage. Likewise, the category "Language Description" might need further refinement depending on the insights during the upcoming survey period.

5.2.2 Level of recommendation

Apart from grouping formats into functional categories, another issue reflected in format white lists and recommendations was the varied means of expressing the extent to which a centre would be ready to accept particular kinds of data and data formats. Centres are willing to go to varying measures in order to ensure data deposition. Some kinds of data are too valuable not to invest the centre's resources in conversion and curation. In most cases, the centre needs to conserve its resources and expects the donor to take care of the easy details, such as the format. For this reason, the centre's interest is not merely binary – "interested" vs. "not interested" – but rather (apart from cases where the data in question is extremely valuable), the scale is minimally composed of three values: recommended, acceptable (can be up-converted with relatively little effort), and deprecated (effectively discouraged – the cost of up-conversion from that format may outweigh the value of the data by far). Note that the very fact that each centre's recommendations may easily differ in this regard, due to traditions of supporting certain formats or local research communities, speaks against an attempt to formulate any sort of specific top-down format recommendations.

How strictly centres need to control incoming deposits depends on the intended further processing. Some centres distribute data sets more or less as they were deposited, with additional standardized metadata added in the deposition process, while other centres want to make sure that all deposited resources comply with requirements at various levels, in order for them to be further enriched, visualized, and/or integrated into a local search engine. The question of whether or not to accept data in non-compliant formats and possibly curate it can be answered by assessing the data value, which is difficult to operationalize, and the curation cost, which is often very hard to estimate for inconsistent and/or legacy data sets. On the other hand, if a repository offers data sets with highly varying characteristics, it becomes very difficult for people wanting to reuse resources to determine which reuse scenarios would be possible for an individual resource. One solution, described in Hedeland (2021), would be to formally define different levels of data maturity, in order to describe linguistic resources as being curated and structured to a certain extent. This would allow depositors to comply with requirements suitable for their research project and users to know what to expect from individual data sets.

5.2.3 Granularity

Another aspect of the design of format recommendations is the amount of detail in the description or the point of discrimination between (sub)types of formats, which also varies greatly in the documents and lists published by CLARIN centres. This question is often related to whether centres process and possibly curate deposited data in order to integrate it into services such as platforms for querying or visualization, since the technical workflows are often designed for specific formats, not for generic formats such as XML[20] or TEI (TEI Consortium, 2021). The same is true for tools and services provided via the infrastructure, and this practical relevance of specific format descriptions became visible in the development of the CLARIN-D WebLicht web service orchestration platform at the German CLARIN centre EKUT (cf. Section 2.1). Since WebLicht uses its own internal format, TCF,[21] various German centres provided converters from their preferred formats to TCF. Users would then upload their data to WebLicht and a suitable converter would be suggested on the basis of the media type. However, it soon turned out that the IANA media type for TEI data (application/tei+xml) was not sufficient to differentiate between, for example, the DTA Base Format for printed texts (DTABf, Haaf, Geyken, and Wiegand, 2015) used at the German CLARIN centre BBAW and the TEI-based ISO 62462:2016 "Transcription of spoken language" (ISO/TC 37/SC 4, 2016) used at the German CLARIN centres HZSK and IDS. Since the respective converters provided by the BBAW and the HZSK were specific to their preferred TEI variants, users' requests for conversion from more or less random TEI customizations to TCF would often fail.

Apart from the two variants of the TEI mentioned above, several other well-documented TEI-based formats are used within CLARIN. These are either tailored to specific research areas, such as Parla-CLARIN (Erjavec et al., 2022) for parliamentary data, CMC-core (Beißwenger and Lüngen, 2020) for computer mediated communication data, or MENOTA (Haugen, 2019) for Nordic medieval texts, or they are locally used variants such as the I5 format (Lüngen and Sperberg-McQueen, 2012) of the IDS centre in Germany, the TEIP5DKCLARIN (Asmussen, 2015) of the Danish consortium, or the TEITOK system (Janssen, 2021) now hosted by the LINDAT centre. These formats are not interchangeable and though they can all be described as "TEI", such a generic description is often insufficient. For the WebLicht use case, a solution was suggested based on required and optional parameters added to the IANA media type by analogy to, for instance,

20 https://www.w3.org/TR/xml/

21 https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

charset for text files (cf. Schmidt, Hedeland, and Jettka, 2017). The parameter "format-variant" was successfully used within WebLicht, and it is a good example of how standardization should follow community practice: at the time of writing, this parameter is not yet part of any official standard for the relevant IANA media types, and any initiative to make it officially recognized must wait until the practice is sufficiently well established within the community.

Similar cases of related but different formats can be found in the CoNLL[22] family and for other TSV-based (tool) formats, which means that this is not a TEI-related problem, but a more general one. And when considering audio and video data with varying codecs, as well as quality-related parameters specified in the recommendations published by CLARIN centres and other organizations, it becomes clear that also for this kind of non-textual data, media type labels are often insufficient for the purpose of an adequate format identification. The PRONOM initiative of The National Archives and the Sustainability of Digital Formats Website of the Library of Congress (cf. Section 5.2) have both found ways of dealing with this very issue. The PRONOM PUID Scheme specification (Brown, 2006, 5), which explains the minting and use of persistent unique identifiers for formats, stresses the importance of granularity decisions and describes how the system differentiates at a fine-grained level:

> The granularity at which separate formats are identified is a crucial feature of the scheme. The PUID identifies formats at the most specific possible level of granularity. For example, the eXtensible Markup Language (XML) is a format which exists in a number of different versions (currently 1.0 and the forthcoming 1.1).

On the other hand, for other features, such as the image compression algorithms of the TIFF 6.0 format (Adobe Developers Association, 1992), no individual PUIDs are issued. In comparison, the Library of Congress issues an ID to the TIFF 6.0 format[23] and also for individual subtypes according to the various compression algorithms. In the context of digital language resources, source data quality is crucial, which was also reflected in the existing recommendations by parameters such as sampling rate and bit depth for audio recordings. Figure 3 shows how this information, encoded as comments in the SIS, discriminates between two entries with different levels of recommendation for the format "WAVE" by the IDS centre. At the time of writing, there is no final solution to these questions for the SIS,

22 CoNLL formats have been born in the context of shared tasks of the SIGNLL Conference on Computational Natural Language Learning https://www.conll.org/. The most popular of them is CoNLL-U (https://universaldependencies.org/format.html), with a template for extensions; versions of CoNLL-U addressing word lattices and anaphora resolution have also been proposed.
23 https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml

but as in the case of the functional domains (cf. Section 5.2.1), the CSC intends to base decisions regarding granularity of format descriptions on the practical use by service providers and users of the infrastructure.

5.3 Standards Information System: Data model

Figure 1 below presents the addition of format recommendation information to the (simplified) data model of the earlier version of the SIS. What can be seen in the diagram is that formats, while in most cases defined by published standards, are a class of their own, with information that is in many cases independent from standards, such as the recommended file extension or the recommended MIME type – information items that have proven to be of use to CLARIN developers.

For the purpose of aggregating and visualizing deposition format recommendations, we consider a single instance of recommendation as a qualified link between a triple: {Format, Domain, Centre}, where the former two are combined in what is basically a Cartesian product dubbed "Relativized format" – that is, a format that realizes a function described by the given domain. For example, the following recommendation: {FLAC, Audiovisual Source Language Data, IDS}, qualified as "acceptable", expresses the fact that the IDS declares it will accept depositions in the FLAC format for data belonging to the domain "Audiovisual Source Language Data".[24]

5.4 Workflows for format recommendations

Several workflows have been considered for creating and maintaining format recommendations, depending on what the subparts of the system were assumed to be – for example, while the KPI (Google) spreadsheet was still the locus of format-related information, users of the published system were expected to interact with the spreadsheet via Google Forms. That required third-party add-ons for Google Forms and a lot of data manipulation within the spreadsheet in order to populate the Forms adequately, as well as a non-trivial XML transformation from Forms into the SIS, for visualization. In June 2021, the CSC decided to abandon the KPI spreadsheet and to make the SIS the basis for user workflows. The workflow that is currently envisioned is described below, taking advantage of predefined templates for each depositing centre and of the fact that many formats are already

24 See e.g. https://standards.clarin.eu/sis/views/view-format.xq?id=fFLAC for an implementation.

described within the system. Note that, at this point, the existing format recommendations announced by centres on their home pages must be additionally interpreted in order to be converted into the qualified {Format, Domain, Centre} triples. This in turn means that centre representatives should approach the initial information presented by the SIS with an eye to modifying it to make sure that it fully reflects the given centre's stance. Naturally, the workflow is designed to be applied iteratively, whenever the given centre decides to adjust its recommendations. New centres can also apply it by reusing recommendations from other centres and editing them appropriately.

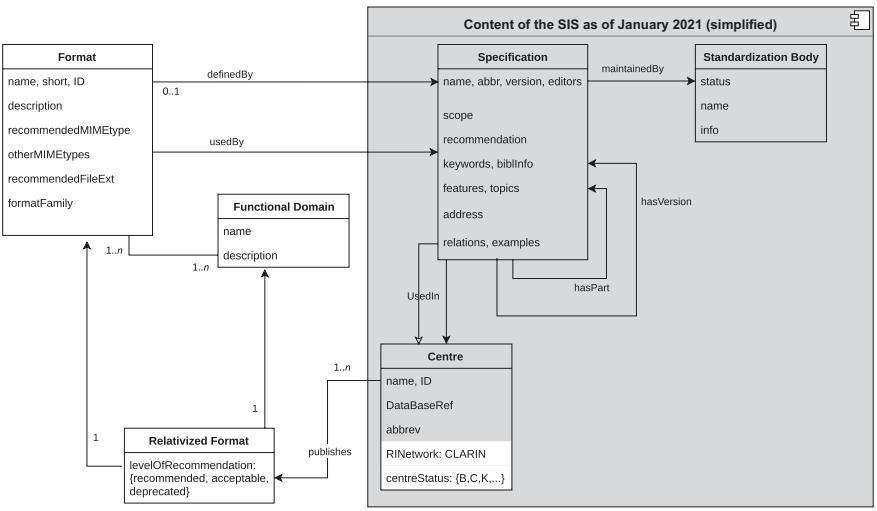

Figure 1: Simplified data model of the SIS; the original parts on bluish background.

The example centre representative (assume that the centre is IDS Mannheim) should start by checking the recommendations encoded for their centre at https://standards.clarin.eu/sis/, either by opening the menu item "Format recommendations" and selecting "IDS" in the first drop-down menu to filter the results, or by opening the IDS-related section from the menu item "Centres". The next step is to verify that the recommendations are correct and complete, and the SIS assists with this by making it possible to sort the data by any column. Figure 2 shows an example screenshot of the filtered sorted recommendations screen, while Figure 3 is a fragment of centre-specific information screen, where additional comments are also shown.

IDS | Select domain ... | Select recommendation ... | Filter | Reset

Search format | Search

Export Table to XML

Format	Clarin Centres	Domain	Recommendation
AIFF	IDS	Audiovisual Source Language Data	acceptable
ALTO	IDS	Text Annotation	acceptable
ANVIL	IDS	Audiovisual Annotation	acceptable
CHAT	IDS	Audiovisual Annotation	deprecated
CHAT-XML	IDS	Audiovisual Annotation	deprecated
CMDI	IDS	Metadata	recommended
Coma	IDS	Metadata	recommended
CSV	IDS	Metadata	acceptable
DC XML	IDS	Metadata	recommended
DGD-XML	IDS	Metadata	recommended
DOCX	IDS	Audiovisual Annotation	deprecated
DOCX	IDS	Metadata	deprecated
DTABf	IDS	Text Annotation	recommended

Figure 2: Example screenshot of format recommendations in the SIS (v. 2.2.0), filtered for "IDS" and sorted alphabetically by format names.

After the potential filtering, recommendations can be exported as XML, to yield a listing similar to the example fragment shown in Figure 4.

This is an editable file that can be modified or extended as necessary, and afterwards submitted to the SIS via GitHub: either by means of a pull request from a forked repository, or by opening the relevant document in the browser and pasting the new content, thus creating a new commit.[25] The commit will be checked for well-formedness and content errors, and eventually uploaded to the live instance of the SIS. If the file is edited with XML-aware software, the underlying schema restricts the options for functional domains and recommendations (they are presented as drop-down lists with glosses for each option).

25 The document relevant for this example resides at https://github.com/clarin-eric/standards/blob/master/SIS/clarin/data/recommendations/IDS-recommendation.xml.

TEISpoken	Audiovisual Annotation	recommended	
plainText	Audiovisual Annotation	deprecated	
Transana	Audiovisual Annotation	deprecated	
TRS	Audiovisual Annotation	acceptable	
AIFF	Audiovisual Source Language Data	acceptable	
FLAC	Audiovisual Source Language Data	acceptable	
M2J ⊕	Audiovisual Source Language Data	acceptable	
MP3	Audiovisual Source Language Data	deprecated	lossy formats should be avoided if possible
MP4	Audiovisual Source Language Data	acceptable	
MPEG-4 AVC	Audiovisual Source Language Data	recommended	25 fps, 1920×1080, constant bit rate
MPEG-1	Audiovisual Source Language Data	acceptable	
MPEG-2	Audiovisual Source Language Data	acceptable	
WAVE	Audiovisual Source Language Data	recommended	PCM-WAV, 48 kHz, 16 bit
WAVE	Audiovisual Source Language Data	acceptable	PCM-WAV with non-recommended parameters (not 48 kHz, 16 bit)

Figure 3: Fragment of the centre-specific information page of the IDS, sorted by domain names, showing example comments that differentiate between seemingly conflicting recommendations.

The live system is cross-linked to predefined GitHub "tickets", which are a way of communicating to the developers and users that something can be added, extended, or fixed. An example of that is shown in Figure 3, where the format "M2J" does not yet have a corresponding information page and the "+" sign indicates that clicking on it will open a GitHub ticket.

6 Summary and outlook

The present chapter provides a glimpse of the work of the CLARIN Standards Committee and locates it within the context of the evolution and maturation of a distributed research infrastructure that needs to establish balance between, on the one hand, the top-down requirements of uniformity and, on the other, the bottom-up tension that stems from freedom of research and the complexity of the target fields of interest. Such a balance contributes to ensuring a satisfactory measure of interoperability among the growing network, and a uniform basis for outreach.

The current picture is one in which a top-down frame of general research principles (FAIR and others) is set over a predefined information structure, which the individual centres can fill in by using shared (and open-ended) taxonomies, in fulfilment of the assessment criteria, but also in order to communicate their profile in practical and uniform terms, both to other centres and to outside users.

Expanding the existing information and adding new centres is simple and transparent, with the contributions guaranteed to be attributable and under version control.

```
<format id="fWave">
   <domain>Audiovisual Source Language Data</domain>
   <level>recommended</level>
   <comment>PCM-WAV, 48 kHz, 16 bit</comment>
</format>
<format id="fWave">
   <domain>Audiovisual Source Language Data</domain>
   <level>acceptable</level>
   <comment>PCM-WAV with non-recommended parameters (not 48 kHz, 16 bit)</comment>
</format>
<format id="fPDFA">
   <domain>Documentation</domain>
   <level>recommended</level>
</format>
<format id="fTextPlain">
   <domain>Documentation</domain>
   <level>recommended</level>
</format>
<format id="fCMDI">
   <domain>Metadata</domain>
   <level>recommended</level>
</format>
```

Figure 4: XML representation of a fragment of format recommendations.

The nearest future for the CSC will consist in ironing out any wrinkles in how the system and the envisioned maintenance workflow function, adding more visualization options, and, in the next step, in looking at the part of the SIS that addresses standards in order to make it as useful in practical terms as the format-related part promises to be. Apart from CLARIN-internal dissemination of information and documentation on the SIS, integration with existing generic initiatives by the Library of Congress and The National Archives (cf. Section 5.2) and the more recently developed FAIRsharing platform[26] is also being considered in order to reach out to users and research infrastructures beyond CLARIN. Furthermore, the SIS could also become valuable as a sound knowledge basis for initiatives targeting interoperability, such as the SSHOC Conversion Hub.[27]

26 https://fairsharing.org/standards/

27 https://conversion-hub.sshopencloud.eu/

Annex A: Major standards-related recommendations in the history of CLARIN

These are standards guidelines that have been circulated as semi- or fully official in the history of CLARIN. This list is most probably incomplete and the ordering is not generally meant to indicate the level of influence or importance, except for the first item, which is important because it was a product of a task force of specialists, and because (as such) it received a lot of attention, and is referenced in many of the other items listed here.

1. *Standards for LRT*, a 2009 document provided on the CLARIN website (at the standards recommendations page, https://www.clarin.eu/content/standard-recommendations) – the most recent version comes from March 2009. This is a relevant document because it is commonly referenced, and because it has been prepared by a committee of experts and representatives of several projects.
2. CLARIN preparatory phase deliverable *D5.C-3: Interoperability and Standards*, edited by Erhard Hinrichs and Iris Vogel. 2010. https://office.clarin.eu/pp/D5C-3.pdf. This document was created in the D-Spin preparatory phase for CLARIN-D and later became the basis for the DFG recommendations for technical and software aspects of corpus creation (cf. 12 in this list).
3. *Standards and Formats*, an overview of recommended CLARIN standards on the CLARIN website: https://www.clarin.eu/content/standards-and-formats
4. *Overview of standard related resources in CLARIN centres*, compiled by Maik Stührenberg in 2014 with input from the CSC: https://trac.clarin.eu/attachment/wiki/StandardsCommittee/Overview_of_Standard-related_resources-2014-08-26-TLA_DK_PL.docx (restricted access).
5. *CLARIN standards guidance* (later renamed *Standards Information System* and deposited at GitHub) hosted at the IDS; see Section 5 of the present chapter.
6. *What standards are recommended by CLARIN?* is a CLARIN website FAQ item: https://www.clarin.eu/faq/what-standards-are-recommended-clarin.
7. *CE-2014-0421 "Relevant data formats"* (https://www.clarin.eu/sites/default/files/CE-2104-0421-relevant-formats.pdf).
8. *"CE-2014-0421-relevant-formats"* (an internal spreadsheet accompanying CE-2014-0421).
9. *DMPTY*, the (experimental) CLARIN-D data management plan wizard (Trippel and Zinn (2015), https://www.clarin-d.net/en/preparation/data-management-plan) refers to the CLARIN-D User guide (cf. 11.) and provides a short (outdated) list of formats.

10. *Format Registry*, a collection of format recommendations mainly resulting from a recent (2015) survey on formats accepted by German CLARIN centres: https://trac.clarin.eu/wiki/FormatRegistry (restricted access).
11. *CLARIN user guide* (in German and English; the former appears to be no longer available in full) https://media.dwds.de/clarin/userguide/text/ (since 2012, but the most recent version is from 2019). An alternative link is: https://www.clarin-d.net/en/language-resources-and-services/user-guide.
12. *DFG Handreichung: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora*, 2nd edition, 2019 (hosted at the DFG website: http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf; there is an unofficial English translation (of the 1st edition) in preparation: Recommendations for Technical Standards and Tools for Building Language Corpora).
13. *Adoption and implementation of standards*, a CLARIN-Plus document authored by Claus Povlsen and Lene Offersgaard (CLARINPLUS-D5.3-7): https://office.clarin.eu/v/CE-2016-0879-CLARINPLUS-D5_3-7-Standards.pdf (referencing the SIS but indirectly also the LRT Standards document).
14. CLARIN Short Guides:
    a. *Standards for text encoding* (May 2009): https://www.clarin.eu/sites/default/files/standards-text-CLARIN-ShortGuide.pdf
    b. *Standards and best practices* (Feb 2009): https://www.clarin.eu/sites/default/files/standards-CLARIN-ShortGuide.pdf
    c. *Web services interoperability* (Feb 2010): https://www.clarin.eu/sites/default/files/ws_interop-CLARIN-ShortGuide.pdf
15. *Interoperability* webpage at https://www.clarin.eu/content/interoperability (maintained by the Interoperability Task Force).

Annex B: Functional domains for deposition formats

This section lists the functional domains that correspond to the most common use scenarios to which data deposited at CLARIN centres may be put; see Section 5.2.1 for the motivation behind some of the choices. The current list is to be found at https://standards.clarin.eu/sis/views/list-domains.xq.

Annotation

- **Audiovisual Annotation**
  Annotations of audiovisual sources, usually including a basic rendering of the spoken content (transcription) and sometimes further annotation.
- **Image Annotation**
  Annotations of image sources.
- **Text Annotation**
  Annotations of textual sources/written text, with the original text included or as stand-off.

Data/resource description

- **Metadata**
  Comprehensive structured information including descriptive, structural, and administrative metadata.
- **Catalogue Metadata**
  Basic structured information for discoverability and general description, to be openly provided for harvesting.
- **Contextual Information**
  Structured information on the communicative event or text and its creators (i.e. participants or authors) relevant for analysis.
- **Documentation**
  Unstructured documentation of the resource and its parts, such as corpus or annotation guidelines.

Databases

- **Language Description**
  Structured or unstructured descriptions of linguistic varieties or phenomena, typological databases, etc.
- **Lexical Resource**
  Structured (item-based) resources for lexical and/or conceptual information on units of language (e.g., wordlists, lexicons, WordNets, etc.)
- **Geodata**
  Information on geographic locations.
- **Statistical Data**
  Data from surveys and tests in numeric formats.

Source data

- **Audiovisual Source Language Data**
  Audio or video recordings providing spoken/multimodal or signed language data for research purposes.
- **Image Source Language Data**
  Digitized images of analogue sources of written language data for research purposes (e.g., facsimiles, scans of handwriting, photos of inscriptions).
- **Textual Source Language Data**
  Written unstructured/plain text or originally structured text (e.g., HTML) without linguistic or other mark-up added for research purposes.
- **Contextual Data**
  Images (photos or drawings) or documents relevant to the communicative event or text but not part of the source language data.

Uncategorized

- **Tool support**
  Tool-related formats required for specific functionality of the tool or reliable reuse of resources (e.g., tagsets, annotation schemes, vocabularies, language models, parameter files, and other specifications or settings)
- **Other**
  Functions not covered by the other domains.

Bibliography

Adobe Developers Association. 1992. TIFF revision 6.0. Technical report, Adobe Systems Incorporated. https://www.awaresystems.be/imaging/tiff/specification/TIFF6.pdf (accessed May 3, 2022).

Asmussen, Jørg. 2015. Text formatting. what an annotated text should look like. Technical report, DK-CLARIN WP 2.1. https://info.clarin.dk/clarin-dk-infrastrukturen/vejledninger/text-format.pdf (accessed May 3, 2022).

Bański, Piotr. 2018. Towards unified CLARIN recommendations for the use of standards: a pilot study on "text formats" (CE-2021-1931), Version 0.2, 2018-05-10. Technical report, CLARIN ERIC. https://hdl.handle.net/11372/DOC-164 (accessed May 3, 2022).

Bański, Piotr, Hanna Hedeland & Dieter Van Uytvanck. 2019. Unified list of standards: next steps forward. Poster presented at the Bazaar, CLARIN Annual Conference 2019, Leipzig, Germany, 30 September-2 October. https://www.clarin.eu/sites/default/files/clarin2019_bazaar_csc.pdf (accessed May 3, 2022).

Beißwenger, Michael & Harald Lüngen. 2020. CMC-core: a schema for the representation of CMC corpora in TEI. *Corpus* 20. https://doi.org/https://doi.org/10.4000/corpus.4553

Broeder, Daan, Menzo Windhouwer, Dieter Van Uytvanck, Thorsten Trippel & Twan Goosen. 2012. CMDI: a component metadata infrastructure. In *Proceedings of LREC-workshop "describing LRs with metadata: Towards flexibility and interoperability in the documentation of LR"*, 1–4. ELRA. http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012MetadataProceedings.pdf (accessed May 3, 2022).

Brown, Adrian. 2006. Digital preservation technical paper 2: The PRONOM unique identifier scheme: A scheme of persistent unique identifiers for representation information (DPTP-02). Digital PreservationTechnical Paper 2, The National Archives. https://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf (accessed May 3, 2022).

Chiarcos, Christian, Christian Fäth & Frank Abromeit. 2020. Annotation interoperability for the Post-ISOCat era. In *International conference on language resources and evaluation (lrec) 12*, 5668–5677. https://aclanthology.org/2020.lrec-1.696.pdf (accessed May 3, 2022).

CLARIN Standards Committee. 2020. Pursuing the elusive KPI: Filling the gaps in centre self-published standards-related information. Presentation given at the CLARIN Annual Conference, in September 2020. https://www.clarin.eu/sites/default/files/Clarin2020_bazaar_CSC_Core_1.pdf (accessed May 3, 2022).

Cooper, Danielle & Rebecca Springer. 2019. Data communities: A new model for supporting STEM data sharing [issue brief]. *DigitalCommons@University of Nebraska – Lincoln*. https://digitalcommons.unl.edu/scholcom/109 (accessed May 3, 2022).

CoreTrustSeal Standards and Certification Board. 2019. CoreTrustSeal trustworthy data repositories requirements 2020–2022. https://doi.org/10.5281/zenodo.3638211

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Dar´gis, Orsolya Ring, Ruben van Heusden, Maarten Marx & Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation* https://doi.org/10.1007/s10579-021-09574-0

Gomes, Luís, Ruben Branco, João Silva & António Branco. 2022. Open and inclusive language processing: Language processing services by PORTULAN to meet the widest needs of CLARIN users. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: deGruyter.

Goosen, Twan, Menzo Windhouwer, Oddrun Ohren, Axel Herold, Thomas Eckart, Matej Ďurčo & Oliver Schonefeld. 2015. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. In *Selected papers from the CLARIN 2014 conference, October 24–25, 2014, Soesterberg, the Netherlands*, 36–53. https://ep.liu.se/ecp/116/004/ecp15116004.pdf (accessed May 3, 2022).

Haaf, Susanne, Alexander Geyken & Frank Wiegand. 2015. The DTA "base format": A TEI subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative* 8. https://doi.org/10.4000/jtei.1114

Hajič, Jan, Eva Hajičová, Barbora Hladká, Jozef Mišutka, Ondřej Košarko & Pavel Straňák. 2022. LINDAT/CLARIAH-CZ: Where we are and where we go. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: deGruyter.

Haugen, Odd Einar, (ed.). 2019. *The Menota handbook: Guidelines for the electronic encoding of medieval Nordic primary sources. Version 3.0*. Bergen: Medieval Nordic Text Archive. http://www.menota.org/handbook.xml (accessed May 3, 2022).

Hedeland, Hanna. 2021. Towards comprehensive definitions of data quality for audiovisual annotated language resources. In *Selected papers from the CLARIN Annual Conference 2020, online, 5–7 October 2020*, 93–103. https://doi.org/10.3384/ecp18011

Hedeland, Hanna & Piotr Bański. 2018. Towards CLARIN recommended formats: a bottom-up approach. poster presented at the Bazaar, CLARIN Annual Conference 2018, Pisa, Italy, 8–10 October. https://www.clarin.eu/sites/default/files/CLARIN2018_Bazaar_Hedeland_Banski.pdf (accessed May 3, 2022).

Hinrichs, Erhard W., Marie Hinrichs & Thomas Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, 25–29. USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/P10-4005 (accessed May 3, 2022).

ISO/TC 37/SC 4. 2016. ISO 24624:2016 Language resource management – Transcription of spoken language. http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338 (accessed May 3, 2022).

Janssen, Maarten. 2021. A corpus with wavesurfer and TEI: Speech and video in TEITOK. In *Text, speech, and dialogue*, 261–268. Cham: Springer International.

Jong, Franciska de, Bente Maegaard, Darja Fišer, Dieter Van Uytvanck & Andreas Witt. 2020. Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. In *International conference on language resources and evaluation (lrec) 12*, 3406–3413. https://aclanthology.org/2020.lrec-1.417 (accessed May 3, 2022).

Jong, Franciska de, Bente Maegaard, Koenraad De Smedt, Darja Fišer & Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and responsible data science using language resources. In *International conference on language resources and evaluation (lrec) 11*, 3259–3264. http://hdl.handle.net/1874/364776 (accessed May 3, 2022).

Lüngen, Harald & C. Michael Sperberg-McQueen. 2012. A TEI P5 document grammar for the IDS text model. *Journal of the Text Encoding Initiative* 3. https://doi.org/10.4000/jtei.508

Maegaard, Bente & Steven Krauwer. 2018. Key performance indicators for CLARIN ERIC (CE-2018-1266), Version 2, 2018-11-07. Technical report, CLARIN ERIC.

Maegaard, Bente & Leon Wessels. 2019. Measuring CLARIN's key performance indicators (CE-2019-1515), version 1.4, 2019-09-24. Technical report, CLARIN ERIC. Draft.

Odijk, Jan. 2016. Towards Interoperability in CLARIN (CE-2016-0845),Version 1, 2016-08-25. techreport Version 1, 2016-08-25, CLARIN ERIC. https://office.clarin.eu/v/CE-2016-0845-towards-interoperability.pdf (accessed May 3, 2022), draft, Distribution NCF.

Olsson, Leif-Jöran. 2017. Federated content search engine v2 (software), CLARINPLUS-D2.9 (CE-2017-1035), 2017-06-09. techreport, CLARIN PLUS. https://office.clarin.eu/v/CE-2017-1035-CLARINPLUS-D2_9.pdf (accessed May 3, 2022), distribution Public.

Pennock, Maureen, Paul Wheatley & Peter May. 2014. Sustainability assessments at the British Library: Formats, frameworks and findings. In *Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6–10, 2014*, 141–148. https://doi.org/11353/10.378694

RDA FAIR Data Maturity Model Working Group. 2020. FAIR data maturity model. specification and guidelines (1.0). techreport, RDA. https://doi.org/10.15497/rda00050

RfII. 2020. The data quality challenge. recommendations for sustainable research in the digital turn. http://www.rfii.de/?p=4203 (accessed May 3, 2022).

Schmidt, Thomas. 2011. A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative* 1, 1–28. https://doi.org/10.4000/jtei.142

Schmidt, Thomas, Hanna Hedeland & Daniel Jettka. 2017. Conversion and annotation web services for spoken language data in CLARIN. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 october 2016*, 113–130. http://www.ep.liu.se/ecp/136/009/ecp17136009.pdf (accessed May 3, 2022).

Schonefeld, Oliver, Thomas Eckart, Thomas Kisler, Christoph Draxler, Kai Zimmer, Matej Ďurčo, Yana Panchenko, Hanna Hedeland, Andre Blessing & Olha Shkaravska. 2014. CLARIN federated content search (CLARIN-FCS) – core specification (CE-2014-0316), version 1.0, 2014-04-07. techreport, CLARIN ERIC. https://svn.clarin.eu/FederatedSearch/docs/CLARIN_FCS_Specification_Core_1_0.docx (accessed May 3, 2022), draft for approval by SCCTC, Distribution SCCTC.

Stührenberg, Maik, Antonina Werthmann & Andreas Witt. 2012. Guidance through the standards jungle for linguistic resources. In *Proceedings of the LREC-12 workshop on collaborative resource development and delivery. Istanbul, Turkey, May 2012*, 9–13. European Language Resources Association (ELRA). https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/4494/file/Stuehrenberg_Werthmann_Witt_Guidance_through_the_standards_jungle_2012.pdf (accessed May 3, 2022).

TEI Consortium. 2021. TEI P5: Guidelines for electronic text encoding and interchange. Technical Report 4.3.0, TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (accessed May 3, 2022).

Thomas, Christian & Frank Wiegand. 2015. Making great work even better: Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th–19th cent.) in CLARIN-D. In *Historical corpora. challenges and perspectives*. Tübingen: Narr. https://www.deutschestextarchiv.de/files/Thomas-Wiegand-2015_Making-Great-Work-Even-Better_CLIP-5_2018-07-05.pdf (accessed May 3, 2022).

Trippel, Thorsten & Claus Zinn. 2015. DMPTY – a wizard for generating data management plans. In *Selected papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*, 71–78. Linköping: Linköping University Electronic Press. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=123&Article_No=6 (accessed May 3, 2022).

Van Uytvanck, Dieter. 2014. Relevant data formats (CE-2014-0421), Version 1, 2014-10-17. Technical report, CLARIN ERIC. https://www.clarin.eu/sites/default/files/CE-2104-0421-relevant-formats.pdf (accessed May 3, 2022), draft, Distribution SCCTC.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, 160018–. https://doi.org/10.1038/sdata.2016.18

Windhouwer, Menzo & Twan Goosen. 2022. Component Metadata Infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: deGruyter.

Wittenburg, Peter, Daan Broeder, Bertrand Gaiffe, Maria Gavrilidou, Erhard Hinrichs, Lothar Lemnitzer, Dieter Van Uytvanck & Andreas Witt. 2008. Metadata Infrastructure for Language Resources and Technology, (CLARIN-2008-5), D2.4, Version 5. Technical report, CLARIN. https://www.clarin.eu/sites/default/files/wg2-4-metadata-doc-v5.pdf (accessed May 3, 2022).

Wittenburg, Peter, Dieter Van Uytvanck, Thomas Zastrow & Lene Offersgaard. 2020. CLARIN centre types (CE-2012-0037), version 0.8, 2020-02-18. techreport Version 0.8, 2020-02-18, CLARIN ERIC. http://hdl.handle.net/11372/DOC-77 (accessed May 3, 2022), for approval by SCCTC, Distribution SCCTC, CAC, BoD.

Wittenburg, Peter, Dieter Van Uytvanck, Thomas Zastrow, Pavel Straňák, Daan Broeder, Florian Schiel, Volker Boehlke, Uwe Reichel & Lene Offersgaard. 2019. CLARIN B centre checklist (CE-2013-0095), version 7.3.1, 2019-09-30. Technical report, CLARIN ERIC. http://hdl.handle.net/11372/DOC-78 (accessed May 3, 2022).

Zinn, Claus & Emanuel Dima. 2022. The CLARIN Language Resource Switchboard: Current state, impact, and future roadmap. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: deGruyter.