Check where data-deposition centres point for format recommendations #14

bansp commented on Mar 30, 2021 • edited Member

Date stamp: 6 October 2024

0. Introduction

0.1. About this very note

This lead comment of ticket #14 in the repository of the Standards Information System (SIS) is part of the release packages of CLARIN recommendations for data-deposition formats, and as such it gets updated at least once per release. The minimal amount of information expected to change cyclically is an update of the date string, after having confirmed that the information is current. Ideally, we are hoping for the centres listed in sections 1-4 below to gravitate towards section 5, which lists centres that maintain information in the SIS.

The information contained herein is meant to assist various bodies in the broadly conceived CLARIN governance (notably the Standards and Interoperability Committee, the BoD and the NCF, the Assessment Committee, and especially the Technical Centres Committee) in their review- and decision-making processes.

This page is located at `https://github.com/clarin-eric/standards/issues/14` . Please post remarks, updates and/or corrections in the comments section at the bottom.

0.2. General introduction (a.k.a. "Why bother")

CLARIN centres often offer deposition services.
B-centres that offer such services are obligated (this is an (re-)assessment precondition, formulated as part of the CoreTrustSeal requirements) to publish explicit information about data formats that they recommend for depositions. For non-B-centres, this is not a requirement, but it is not uncommon, depending on the centre's profile and infrastructure. That obligation/practice has been encoded in one of CLARIN's Key Performance Indicators, using the following measurement: "percentage of centres offering repository services that have published an overview of formats that can be processed in their repository". (Thus the KPI measurement encompasses centres with deposition services, whereas the CTS requirement pertains to B-centres with deposition services, i.e., a subset of the KPI target group; for more details, quotes and references, see section 4 of chapter "Standards in CLARIN", by Piotr Bański and Hanna Hedeland (2022), in the CLARIN Book.)

Before the SIS took wing, the requirement / good practice of publishing explicit information on recommended formats had been addressed in the following ways:

1. publishing the information somewhere at the centre (or consortium) homepage;
2. not publishing that information and instead directing users to by now obsolete sets of recommendations (called "external guidelines" in what follows) that are far too general to

represent the given centre's research profile;
3. using a mixture of the above approaches.

There was/is also a fourth group, consisting of centres with deposition services that wouldn't publish such information at all, not even as a link. It is hoped that this group is going to dissipate soon, especially thanks to the recent initiative by the Technical Centres Committee, inviting centres to deposit the relevant information in the SIS.

This very ticket is devoted mainly to collecting information on centres with depositing services that point to external sources of information on format recommendations, especially if those sources are not very informative. In other words, we are looking mainly at groups 2 and 3.

The reason for listing this information is:

- to have general info on the "trends" in handling the assessment requirement,
- to produce information potentially useful at least to the Technical Centres Committee and to the Assessment Committee,
- to identify the specific targets to see if something needs/may be done about them in order to tighten the information system,
- to assist centres in publishing centre-specific data-deposition format recommendations.

Eventually, the format recommendations are expected to be collected in the Standards Information System. It is possible for centres to store that information in the SIS, and to present it to users with a dedicated link, such as:

https://standards.clarin.eu/sis/views/view-centre.xq?id=IDS

It is also possible to retrieve the information *from* the SIS already pre-structured, as XML, to be styled according to the given centre's guidelines and publish on that centre's pages, this way avoiding the chore of maintaining two separate sets of data (for more on that, see the API section of the SIS).

0.3. Methodology

The primary resource assumed for this task is the CLARIN centre registry at https://centres.clarin.eu/ .

Two secondary resources are:

- the list of depositing centres at https://www.clarin.eu/content/depositing-services
- the list of B-centres at https://www.clarin.eu/content/certified-b-centres

The secondary resources appear to depend on the CLARIN centre registry and a degree of hand-crafting (and therefore a potential update lag) may probably be assumed of the depositing-services page.

A tertiary resource is the list provided by the SIS, at https://standards.clarin.eu/sis/views/list-centres.xq . While one might be tempted to assume that that list should be at least semi-automatically derived from the centre registry, it actually provides a small potential layer of indirection, at least in two aspects: firstly, we allow centres to override the shorthand handles that are listed in the registry (and thus, for example, at the centre's request, "CLARINSI" is listed as "CLARIN.SI") and, secondly, we are prepared for a degree of "ontological" or organisational variability in the case of centres that act as nodes in more than a single research-infrastructure network. In short: centres can influence their listings in the SIS in various ways, independent of the CLARIN registry.

The B-centre status is conditioned upon a successful round of certification, managed internally by the Assessment Committee, and externally by a certification authority, currently the CTS and, in the future, also nestor. (Note, incidentally, that full-fledged methodology would probably ideally start from the CTS database as the primary source, but do forgive us for not trying to shoot gnats with rockets -- the amount of time allocated to this already extensive exercise should be reasonable). The CLARIN registry has various status strings for centres that wish to achieve the B-status, whether for the first time or having lost it and preparing for another certification round -- it is, as of June 2024, "Aiming for B", "Aiming for B.", "aiming for B" (kudos for consistency) but also "none" or "Certification expired, renewal planned". In the present note, all such centres, together with "regular" C-centres, are going to be treated as "Non-B centres". Note that, especially for centres tagged as "none" in the registry, some degree of network-internal knowledge is going to be necessary for stating which of the centres are temporarily not B only because they are getting, or preparing to get, re-certified. There's no guarantee that that knowledge is perfect, so this is a weak point in the methodology.

The tables below are constructed by scrolling along the CLARIN registry and the SIS list in parallel, taking into account (a) B-centres and (b) non-B centres with deposition services that are known as such (note: this is a weak spot, some may escape, and the secondary CLARIN list is not trusted fully). In the process, the SIS list is updated wrt the CLARIN registry, and the result is sorted into the categories provided in the sections that follow. Doubts that arise wrt to the nature of individual centres are usually signalled by GitHub issues that use the "centre data" label.

0.4. Terminology

When, in what follows, a centre is said to "point to external guidelines", those guidelines are in too many cases general, top-down, coarse-grained standards recommendations that were formulated well over 10 years ago and were meant for a purpose different than informing users about centre-particular recommendations on what kinds of data the given centre can handle or is interested in handling. While such pointers are surely provided in good faith, they can at best be considered tricks for passing CTS certification. Otherwise, for practical purposes, *they don't get the thing done*.

Another piece of terminology: "listed in the SIS" vs "curated": some of the content in the SIS comes from rather quick import of information that was structured rather differently back when the Standards Committee worked with spreadsheets. A lot of interpretation happened on the way between spreadsheets and the SIS, justified by the hope that the centres would quickly want to fix that if they were not happy with the outcome. It later turned out that we were a non-tiny bit too hopeful about that. The "legacy" listings, not approved by the particular centres, are accompanied by a warning in red. When, on the other hand, a centre decides to hold an [inputhon](#) and submits the result to the SIS, such recommendations are considered curated and the red warning is replaced with the name(s) of the curator(s) (see Section 5 for examples).

* * *

What follows is information on how the particular CLARIN centres publish format recommendations or how they do *not* publish that info while nevertheless trying to satisfy the CLARIN-internal as well as CTS-imposed requirements. Note the date stamp at the top of this note and please do not hesitate to let us know (ideally: in the comments below) if you see that some info can/should be updated or fixed.

1. Centres that point solely to external guidelines

This section lists centres that do not provide information specific to their research profiles but rather point to general and coarse-grained information provided by CLARIN quite a while ago, in most cases in 2009 (the "LRT Standards" document, which simply doesn't help and mentions obsolete standards).

Note: it is good *not* to be mentioned in this section.

Methodological note: it is possible that the centres below also provide their own recommendations or even point at the SIS for that -- and that that mention has been overlooked (or it has been added after the date stamp in the table below). Gathering data for this ticket has shown that some such information can be located in non-obvious places (that is very rare, but it has happened). In such cases, a question would arise as to how effective is a hidden pointer to the SIS or a hidden table of recommendations, and how that corresponds to the need of satisfying a KPI or a CTS/assessment requirement.

1.1. B-centres

Recall that B-centres are obligated (by CTS Requirement 8) to provide explicit information on what formats they are willing to process in the deposition process. The centres below instead point to general and at least partly obsolete guidelines. Amending this situation is at this point easy: deposit the relevant information directly in the SIS -- and then point to that description.

Centre	LastChecked	LinkTarget
CLARIN-LV	05-10-2024	http://www.clarin.eu/sites/default/files/Standards%20for%20LR[illegible] v6.pdf
CLARIN-PL1	05-10-2024	http://www.clarin.eu/sites/default/files/Standards%20for%20LR[illegible] v6.pdf

Centre	LastChecked	LinkTarget
ILC4CLARIN	05-10-2024	http://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
LINDAT	05-10-2024	http://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf

1.2. Non-B centres

These centres are not obligated to explicitly publish information about what formats they recommend for deposition. However, that is both useful for the users themselves, and also crucial for satisfying the relevant CLARIN KPI. Also, these centres are listed as aiming for the "B" status (click on "Type status" to sort them at the top), so at some point they will need to undergo CTS assessment -- why not be proactive in this respect.

Centre	LastChecked	LinkTarget
CLARIN-LT	05-10-2024	http://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
ERCC	05-10-2024	http://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
SADiLaR	05-10-2024	http://www.clarin.eu/recommendations https://archive.mpi.nl/accepted-file-formats

- SADiLaR points to mpi.nl recommendations, which is definitely more helpful than just pointing to the "LRT standards" document, although a question arises as to whether the recommendations pointed to are actually the centre's own recommendations, given that the centre doesn't have any control over the referenced list. But that is for the CTS to assess. Pointing to the "recommendations" at clarin.eu (after that page has been changed) is a big plus.

2. Centres that point to external guidelines in addition to publishing own information locally

Just like centres listed in section 4 below, those listed in this section fulfil the CTS requirements by publishing explicit requirements concerning formats in which data can be deposited with them. **They have done a splendid job.** The place where their won recommendations are published are listed in the last column of the table below.

The focus of this note is to see where centres point for external information, and, in particular, to catalogue the (let's call them) suboptimal places where users are directed, so that something can be done about that. There is nothing wrong in pointing to an external source in addition to the centre's own recommendations published on the centre's own pages, especially if the external resource brings in some extra value (see ACDH-ARCHE for an example, pointing to Archeology Data Service recommendations). On the other hand, it is not so good to point to obsolete, unhelpful or misleading documents, such as the "LRT standards" PDF.

Thus, the role of this section is basically informative, though with a request directed at the centres enumerated below, to consider sharing at least the positive recommendations in the SIS, in order to enable aggregation of this information and publishing it for the benefit of the community.

(There seems to be no need to split the centres listed here into B- and non-B-. Centres that point to external info AND at the same time maintain their recommendations in the SIS, such as CLARIN-CH, are only listed in Section 5 below, and potential "suboptimal" external links are enumerated under that last table.)

Centre	LastChecked	LinkTarget
ACDH-ARCHE	05-10-2024	a.o. https://www.clarin.eu/content/standard-recommendations
BBAW	05-10-2024	https://www.clarin-d.net/en/language-resources-and-services/user-guide
CLARIN.SI	05-10-2024	http://www.clarin.eu/sites/default/files/Standards%20for%20LRT v6.pdf https://www.clarin.eu/content/standards-and-formats
DH-REP	05-10-2024	https://files.dnb.de/nestor/materialien/nestor_mat_08_eng.pdf
ORTOLANG	05-10-2024	https://www.clarin.eu/content/standard-recommendations
TGrep	05-10-2024	https://files.dnb.de/nestor/materialien/nestor_mat_08_eng.pdf
UdS	05-10-2024	http://www.clarin.eu/recommendations

- CLARIN.SI provides its own info in an exemplary way; it has a "see also" section with links but that section is only for the explorers, because all the relevant data are served on a silver plate. Note: the links to the old CSC spreadsheets are probably an overkill (the info is extremely obsolete) -- a link to the SIS (at least to the general recommendations, if not to the listing of the SI consortium) would be appreciated.
- TGrep recommendations are not easy to locate. That can be considered a usability or user-friendliness issue.

3. Centres that neither point anywhere nor publish their own explicit information

This set of centres should be empty, *unless the centre does not offer deposition services* (in which case, it shouldn't be listed here, so... this set should be empty). Please note that rather than amending the existing lack of recommendations on their own home pages, the best course of action for these centres may be to deposit the information directly in the SIS, and then point to that listing. You do one inputhon and Bob's your... list.

3.1. B-centres

"Absence of evidence is not evidence of absence", and it might be that the centres here do publish their own recommendations, in a non-obvious corner of their homepages. Please feel very welcome to post a comment below if you are able to share info on that. Note also that if the info is hidden then it's not really easily available to the depositing users, and ensuring availability of the information is part of the reason for this entire exercise.

Centre	LastChecked	Comment	DepositionPage
CLARIN-IS	05-10-2024	(no info)	https://clarin.is/en/services/

- Note that IS is a relatively fresh B-centre and its position on this page will hopefully change soon (this comment was posted in mid-June 2024)

3.2. Non-B centres

Some of these centres are listed as "aiming for B", some used to be B. All of them indicate that they provide deposition services.

Centre	LastChecked	Comment	DepositionPage
CELR-EKK	05-10-2024	"all data is accepted", via Entu	https://www.keeleressursid.ee/en/services
IMS	05-10-2024	no real recommendations	https://wiki.ims.uni-stuttgart.de/extern/CLARID
IMS (cont.)	15-06-2024	"please contact us"	http://clarin04.ims.uni-stuttgart.de/repo/
MI	05-10-2024	"please contact [us]"	https://meertens.knaw.nl/meertens-collectie/research-data-management/ https://meertens.knaw.nl/en/archive/depositir data_eng_/

- "no real recommendations" for the IMS means the line saying "Is the data in one of the acceptable formats (non-proprietary, text-based) or can it be converted?" -- how is the user to know that? :-(

- The IMS "repo" page was not accessible on 05-10-2024
- MI: the page in Dutch provides more overall information, but neither language version provides a list of formats recommended by the centre

3.3. Special mention

The C-centre CEDIFOR used to "point elsewhere" for data deposition recommendations, but at present it mentions neither CLARIN nor anything concerning data deposition. The maintainer was contacted around June 2024. It is not at all clear that the centre should be mentioned in this ticket (maybe it's not CLARIN, despite the centre registry, or maybe it no longer offers deposition services), so please withhold your judgement.

Centre	LastChecked	Page
CEDIFOR	05-10-2024	https://www.cedifor.de/?s=clarin

4. Centres that only publish their own, local recommendations

Note that this satisfies both the CTS requirement and the KPI calculation (except the KPI calculation performed dynamically by the SIS). Unfortunately, it also ensures a gap in the SIS-derived statistics that might otherwise benefit the entire network. It would be greatly appreciated if at least the data formats *recommended* by the centres could make it into the SIS. The table below includes both B- and C-centres.

Note: Formally, **all these centres have done a splendid job**. Adding their recommendations to the SIS would be a nice bonus to ensure more accurate statistics.

Centre	LastChecked	Linked from / Comment
BAS	05-10-2024	https://clarin.phonetik.uni-muenchen.de/BASRepository/index.ph (choose 'FAQ')
BAS (cont.)	05-10-2024	https://www.bas.uni-muenchen.de/forschung/Bas/BasInfoStandardsTemplateseng.htı
CLARIN-DK	05-10-2024	https://repository.clarin.dk/repository/xmlui/page/faq#what-data formats-are-accepted
CLARIN:EL	05-10-2024	https://www.clarin.gr/en/services/share
CMU	05-10-2024	https://talkbank.org/
COCOON	05-10-2024	https://cocoon.huma-num.fr/exist/crdo/faq.htm?lang=en
DANS	05-10-2024	https://dans.knaw.nl/en/depositing-data-manual/before-depositing_ds/
EKUT	05-10-2024	menu on the main page
IVDNT	05-10-2024	access to the recommendations is not obvious

Centre	LastChecked	Linked from / Comment
LAC	05-10-2024	https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides
MPI-PL	05-10-2024	https://archive.mpi.nl/tla/ + "Help"
TROLLing	05-10-2024	https://site.uit.no/dataverseno/deposit/
ZIM	05-10-2024	https://informationsmodellierung.uni-graz.at/en/about-the-department/research-data-repository-gams/

- ZIM uses GAMS as the repository; it is a repository for digital data from the Humanities / SocSci. The move from ZIM to GAMS required a search for "CLARIN" on the homepage -- there was no obvious path that I could find. It is probably fair to say that the list of the recommended formats is rather coarse-grained. See also issue **ZIM: not sure where to find the format list** #10
- EKUT = TALAR, MPI-PL = TLA
- CLARIN:EL has warned that the paths above may change
- IVDNT provides its own recommendations now (no time stamp in the document -- somewhere, a historian is crying), in a format that should be easy to feed into the SIS. The status of this document is not fully obvious because it does not seem easy to reach the recommendations (or the repository itself) from the homepage of IVDNT. It has crossed my mind that maybe the link to the PDF above is a deep link to an orphaned document (it comes from my old notes). There is a section on CLARIN but it contains no links, and the link to the results of 2023 CTS certification has expired from an Amazon server.

5. Centres that point at their curated recommendations in the SIS

This is where all (or most of) the centres listed above should ideally end up -- what is needed for them is to maintain the information served by the SIS and explicitly link to it. ("Ideally" from the point of view of contributing to the aggregated information; note that for centres in sections 2 and 4, this is a matter of willingness and sparing the time; they are otherwise fine from the point of view of certification and KPI calculation done by hand, rather than in the SIS).

Centre	LastChecked	SourcePage
CLARIN-CH	04-10-2024	https://clarin-ch.ch/documentation-platform/standard-dat formats
CLARINO_Bergen	05-10-2024	https://repo.clarino.uib.no/xmlui/page/faq#what-submissi do-you-accept
FIN-CLARIN	05-10-2024	https://www.kielipankki.fi/tuki/tekninen-muoto/
IDS	05-10-2024	https://repos.ids-mannheim.de/reposdescription.html
OTA	05-10-2024	http://www.clarin.eu/sites/default/files/Standards%20for% v6.pdf

Centre	LastChecked	SourcePage
PORTULAN	05-10-2024	https://portulanclarin.net/usage/#how
SAW	05-10-2024	https://repo.data.saw-leipzig.de/depositing/en
Språkbanken	05-10-2024	https://repo.spraakbanken.gu.se/xmlui/page/faq#what-submissions-do-you-accept

6. Conclusions

6.1. One conclusion that should be drawn from the picture above that the FAQ contained in the LINDAT customisation of DSpace (the deposition system that unifies many repositories, currently) should no longer point users at the LRT PDF but

- by default, reference e.g. https://www.clarin.eu/content/standard-recommendations
- but also encourage the individual centres to customise the link and point directly at the centre's listing in the SIS.

As of October 2024, this is being handled in issue #22 .

6.2. Following up on the above, the default landing page (`content/standard-recommendations`) should, at the top, point at the combined recommendations in the SIS. (that got handled on 17-06-2024)

6.3. Centres which provide their own extensive recommendations will hopefully be willing to share at least their recommended (as opposed to accepted and discouraged) formats, so that (a) the KPI can be properly calculated in the SIS, and (b) so that the statistics of popular formats are not skewed due to the lack of data coming from those centres. This can only be a matter of argumentation and appeal to the "common good" vs. the various restrictions on those centres (time being the most commonly cited one).