"CLARIN-compatible"

Introduction
This document describes the meaning of the term "CLARIN-compatible" when applied to resources (data and software). The focus is on the situation in the Netherlands.

Which resources?
"CLARIN-compatibility" applies to all "language" resources (data and software),

This includes, inter alia, text corpora (with all kinds of annotations), lexicons, databases containing linguistic data, typological databases, audio and video containing speech (speech technology can be used to disclose such data), software specifically targeting these kinds of data, software for making annotations, visualisation software, automatic parsers, PoS-taggers, lemmatizers, spelling normalisation software, chunkers, etc. etc.

CLARIN-compatible
"CLARIN-compatible" includes many aspects that will be briefly described below.

A number of these aspects must be dealt with by the data producer (1-5) as described below:

1. There must be metadata for all resources (data and software).These metadata must be compatible with the Component-based Metadata Infrastructure (CMDI). CMDI offers flexible opportunities to make metadata descriptions, if needed adapted to the requirements specific to a particular resource. A resource is described using metadata according to a metadata profile. Metadata profiles for many common resource types are available and these can be reused, but CMDI also offers the opportunity and the tools to create new profiles if a specific resource or resource type would require that. A profile consists of components. A component consists of metadata elements and/or (recursively) of components again. CMDI offers opportunities to select existing components and to create new components. Finally, CMDI offers a metadata editor (ARBIL) to create CMDI-compatible metadata.
    a. More information on CMDI: http://www.clarin.eu/cmdi (if you want to edit you must login (use your own institute account or if that is not possible create an account on the www.clarin.eu website).
    b. CLARIN-NL regularly organises tutorials and workshops to clarify the CMDI principles and background, to gain hands-on experience, and to discuss problems that may have been encountered.
    c. One can always contact the CLARIN-NL Helpdesk for technical questions on CMDI (and other subjects).
2. Data must be represented in one of a limited number of data formats. The list of data formats and their status in CLARIN can be found on http://www.clarin.eu/recommendations or via the CLARIN-NL website http://www.clarin.nl/node/128 . If, for a particular resource, none of these formats would be suited, contact the CLARIN-NL helpdesk for advice.

3. All data categories that occur in the resource itself or in its metadata must be mapped to a data category in ISOCAT (or in other CLARIN-supported data category registries[1]). In certain cases entries must be created in the RELCAT registry (that is still under development).
    a. Information on ISOCAT: http://www.isocat.org
    b. CLARIN-NL regularly organises tutorials and workshops on this topic as well.
    c. Contact the Helpdesk if you have questions or problems
4. Many data are only really useful if one can search in them efficiently. Data in text formats (e.g. XML) generally are not suited for efficient search (computers are very slow). For this reason it is useful to make available a "live version" of the data (via a CLARIN-centre) that offers efficient search options, e.g. because indexes have been added, special database formats are used, etc. etc. and a search interface is offered (e.g. as a web service and possibly also as a (web) application). The search interface offered must be compatible with the CLARIN federated search requirements (under development). Contact the CLARIN Helpdesk for additional information.
5. The form of software must be fixed in an early stage of development. Two forms are preferred: an *application*, i.e. software with a well-defined user interface (e.g. web based) specifically tuned to the targeted user group; or a *web service*, i.e. software with a well-defined interface to other software and operating on the World Wide Web. For applications it is recommended to implement the core functionality as a web service, and to define the user interface in terms of functions and data structures offered by the web service. There are different protocols for web services: SOAP is the preferred protocol but RESTful services are also allowed. The CLAM system[2] enables one to create a (RESTful) web service for an existing piece of software in an easy way. It is recommended to always contact the Helpdesk for advice on incorporating software in CLARIN.

All resources must be made available via a CLARIN centre. The CLARIN centres in the Netherlands are: Meertens Institute, INL, MPI, DANS and Huygens ING. See http://www.clarin.nl/node/130 for more information.

Other tasks must be carried out by CLARIN centres, of course in close cooperation with the data producer. The CLARIN-centres must

a. Make available the metadata of every resource for metadata harvesting (OAI-PMH protocol),
b. Assign Persistent IDentifiers (PIDs) to every resource (or even parts of it) and to all metadata
c. Make each resource itself available via the CLARIN infrastructure (original data plus, where appropriate, a 'live version').
d. Ensure the long term preservation of the resources

CLARIN-NL is prepared, if there is a need, to set up tutorials on CLARIN-compatibility or specific topics related to it.

---

[1] For example, the registry for country codes ISO3166: http://en.wikipedia.org/wiki/Iso_3166
[2] http://proycon.github.io/clam/

We are well aware that the notion of "CLARIN-compatibility" is continuously developing. In the area of research, which, by definition, must be innovative, new types of resources will come into existence and the existing standards and best practices will not necessarily be suited to such new resource types. We request the researchers or data producers to notify the Helpdesk and to cooperate to the further development of existing standards or to contribute to the development of completely new standards and best practices.

CLARIN-NL Helpdesk:
URL: http://www.clarin.nl/node/134
E-mail: mailto:helpdesk@clarin.nl