Introducing CKAN:
the WP5 collection registration system

By: Willem Melder, Eva Baaren, Liliana Melgar

The "media studies" work package of the CLARIAH project attempts to "plug" five tools developed in different projects to the CLARIAH infrastructure. These "tools" are currently available to media scholars and oral historians, where they can use different collections in the audio-visual domain and closely related collections, such as newspapers[1].

The challenged posed to ICT experts by the media experts in the CLARIAH project was how to be able to use the different collections that were available in each tool in an interchangeable way between all tools. That is, how to use, for example, the collection of oral history interviews that was in the tool "Verteld Verleden" with the tool (and functionalities) of AVResearcherXL. The experts tackled this challenge by proposing the separation of data and functionalities, focusing on two aspects:

(1) the development of the functionalities in a modular approach, offering both "components" and "recipes" (the old tools but built via the combination of different components), and
(2) the identification and description of all datasets or collections available in each tool, and the connection of these datasets to the functionalities offered in the media suite

For the identification and description of the datasets, a system was proposed in which the datasets could be registered. Registering and depositing datasets is a common practice in the research domain, where researchers are requested to make their research data available (e.g., using DANS). There are a few examples of collection registration in other domains, for example in the scientific data sector (B2FIND[2]), or in the linguistic domain (CLAPOP[3], TLA/FLAT[4]).

At WP5, the adopted platform for collection registration is CKAN. It is a software solution that provides the tools to "streamline publishing, sharing, finding and using data,"[5] making it accessible. CKAN is initially focused on making open collections all over the world easily discovered and accessed. However, it can also be used for collections that are not open, as a way to make them discoverable.

[1] http://www.clariah.nl/werkpakketten/focusgebieden/media-studies
[2] B2FIND: http://b2find.eudat.eu/dataset
[3] CLAPOP (CLARIN): http://dev.clarin.nl/clarin-data-list-fs
[4] *http://www.lrec-conf.org/proceedings/lrec2016/pdf/476_Paper.pdf*
[5] http://ckan.org

The current instance of CKAN implemented for the CLARIAH media studies track includes 38 datasets, registered on behalf of the collection owners. For now, organisations included are represented by WP5 people, but involvement of digital cultural heritage representants is definitely of interest, so that improvements of the data and data curation is more distributed and persistent within the near future. Organisations can register datasets and add data. Registration with an open licence is possible, but private access is also an option.

Datasets that are registered can be searched by using tags, organizations, or groups. Groups can be used to organise datasets in relation to a tool or a project in which they could be originally found. Each dataset is described using basic descriptive and administrative metadata. The data itself can be added to each registered collection using different formats (e.g., XML, JSON, RDF). Current work includes: the identification of new relevant datasets to be added, the creation of guidelines that can be used by non-ICT experts willing to add their datasets, and community engagement strategies.

CKAN is aimed at data publishers (national and regional memory institutions and organizations) wanting to make their data open and available. It is also aimed at individual researchers, who can register their own datasets. However, there are problematic issues from the perspective of researchers and collection owners (as well as technical issues) that still need to be discussed. Opening the discussion and sharing ideas will be the main aim of the presentation of CKAN at the CLARIAH tech-day!