Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3457–3462
Marseille, 11–16 May 2020

Digital Language Infrastructures – Documenting Language Actors

Verena Lyding[1], Alexander König[1,2], Monica Pretti[1]
[1] Institute for Applied Linguistics, Eurac Research, Bolzano, Italy
[2] CLARIN ERIC, the Netherlands

verena.lyding@eurac.edu, alex@clarin.eu, pretti.monica@gmail.com

Abstract

The major European language infrastructure initiatives like CLARIN (Hinrichs and Krauwer, 2014), DARIAH (Edmond et al., 2017) or Europeana (Europeana Foundation, 2015) have been built by focusing in the first place on institutions of larger scale, like specialized research departments and larger official units like national libraries, etc. However, besides these principal players also a large number of smaller language actors could contribute to and benefit from language infrastructures. Especially since these smaller institutions, like local libraries, archives and publishers, often collect, manage and host language resources of particular value for their geographical and cultural region, it seems highly relevant to find ways of engaging and connecting them to existing European infrastructure initiatives. In this article, we first highlight the need for reaching out to smaller local language actors and discuss challenges related to this ambition. Then we present the first step in how this objective was approached within a local language infrastructure project, namely by means of a structured documentation of the local language actors landscape in South Tyrol. We describe how the documentation efforts were structured and organized, and what tool we have set up to distribute the collected data online, by adapting existing CLARIN solutions.

Keywords: local infrastructure, language actors, institution catalog

1. Introduction

Existing digital language infrastructures like CLARIN (Hinrichs and Krauwer, 2014), DARIAH (Edmond et al., 2017) or Europeana (Europeana Foundation, 2015) have been built by focusing in the first place on well-established and specialized research institutions and larger official units like national libraries, major European museums etc. This is sensible as it allowed to gather and develop best practices and joint solutions among established centers of expertise to "create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools"[1]. However, besides these principal players and some smaller actors, an even larger number of local language institutions could contribute to and benefit from language infrastructures. This is especially true at this point in time, where the basics for creating language infrastructures are on a good way. As these smaller institutions, like local libraries, archives and publishers, often collect, manage and host language resources of particular value for their geographical and cultural region, it seems particularly relevant to find ways of engaging and connecting them to existing European infrastructure initiatives.

In this paper, we highlight the need to reach out to smaller language actors on a local scope and report on a concrete effort we have made to tackle this need within the context of the DI-ÖSS project[2]. The project aims at taking the first steps towards creating a local language infrastructure for South Tyrol, by bringing together relevant actors of various language institutions and organizations on the local level and by transferring and applying best practices and standards of European initiatives to the local context (see Section 2).

In this paper, we focus on one specialized subtask of DI-ÖSS that is concerned with the systematic screening and documentation of the ecosystem of language actors in South Tyrol. This task strives to involve local target groups, that are oftentimes not linked to bigger infrastructures yet, by actively approaching them and building a concise documentation of their data, services and needs—in terms of workflows and target groups. More specifically, we describe the process of creating a concise documentation of the local ecosystem of language institutions while exploring ways of formalizing the collected information and making it available for consultation online. By doing so, we aim at facilitating information gain about local language actors: *Who is around in the South Tyrolean language context? What are these actors doing with language data and which services are they offering?* And, *where can these actors be found and contacted?* By providing free access to this data in a structured way (see Section 4.3), we aim at fostering collaboration opportunities among institutions.

2. The DI-ÖSS Project

The DI-ÖSS project, running from 2017 to 2020, approaches the overall challenge of making the first steps in growing a digital language infrastructure between various local language institutions by means of implementing prototypical use cases to exploit synergies between the institutions (Lyding et al., 2018).

In fact, the project was initiated with the assumption that a digital language infrastructure could benefit any organization dealing with language data paired with the observation that smaller local institutions are however less involved yet. Given that a lot of knowledge and data about local language and cultural heritage is situated in smaller institutions on the local level and given that these institutions are often less connected to bigger research initiatives, it seems relevant to find ways to actively approach and involve these smaller players.

[1] https://www.clarin.eu/content/clarin-in-a-nutshell

[2] Digital Infrastructure for the Ecosystem of South Tyrolean Language Data and Services: `http://www.eurac.edu/en/research/projects/Pages/projectdetail4262.aspx`

In pursuing the project objectives we have encountered four main challenges, which are related to different characteristics particular to smaller local language actors:

1. Local actors are oftentimes not aware of bigger infrastructure initiatives.
2. Smaller actors often lack the required methodological knowledge, technical skills or human resources for addressing meta-tasks that go beyond the daily duties of their businesses.
3. Smaller local actors, which usually have little or no experience in infrastructure initiatives often encounter difficulties to anticipate the added value of their involvement.
4. The local language ecosystem and the characteristics of individual language institutions and organizations are not transparent and information about them is not openly accessible in a centralized place.

In the following section we will present an approach for addressing the fourth challenge, the need for the systematic documentation of language institutions.

3. Objectives and Overall Approach

As discussed in the previous section, gaining a comprehensive overview of the existing language actors and their role in the local context, i.e. a current-state depiction of each actor's data and data management practices within the local language ecosystem, is of fundamental importance to:

1. Understand the overall local language landscape, and
2. the current situation of the individual institutions and their related demands.

The second point is needed in order to be able to extrapolate from the selected use cases and to identify follow-up opportunities for a wide-reaching infrastructure on the local scope, while the first point allows to gain an overview of the local situation and to identify individual actors.
We therefore claim that digital language infrastructure initiatives should not only be concerned with documenting and linking language data and tools, but also with defining systematic procedures for the documentation of language actors, their organizational structure, functions, resources and needs. We also claim that a common and publicly accessible repository for the documentation of language actors should be set up and we present our approach for tackling this task.
Our approach, within the DI-ÖSS project, aims at mapping out relations and possible interactions between the chosen set of language institutions in order to realize a coherent infrastructural framework of multiple connections between them. It follows a bottom-up (vs top-down) logic which turned to the language actors and, learned about their situation and needs in a first step and documented this situation, with the aim to actively involve them and respond to their needs in a second step. In the remainder of this paper, we will focus on the first step, the bottom-up informed documentation of language actors in South Tyrol.

4. Documenting Language Actors

The documentation process of the local language actors is approached in three steps:

1. Identifying the actors and establishing documentation categories,
2. collecting and organizing the data, and
3. formalizing and distributing them.

The following sections discuss the three working phases in further detail.

4.1. Identifying Actors and Documentation Categories

This primary stage of the documentation process is threefold: firstly, it consists in identifying relevant establishments in the local context, compiling a comprehensive overview of such actors, and categorizing them into self-contained yet interrelated clusters. Secondly, it designates the shortlist of documentation criteria for portraying the chosen institutions in adequate detail and, thirdly, it entails the actual selection of project-apt candidates.
In the first place, an enumeration comprising a total of around 200 establishments was drawn up. In line with the project objectives, the scope for selecting institutions was confined in two ways:

1. The geographical area – the Autonomous Province of Bolzano/Bozen in northern Italy, and
2. the type of institutions – organizations primarily dealing with language data and services.

The list included profile and contact information about each identified establishment as well as a seven-category clustering based on institution types (see Figure 1 for a percentage representation of the preliminary institution classification). The seven institution types are: archives, libraries, online media, catalogs, cultural institutes, publishing houses and journals. These classes were determined in a bottom-up manner, i.e. by observing and abstracting which organizations showed similarities to or shared common ground with one another given their predominant area(s) of interest and competence. While representing a valuable project output in itself[3], such categorization allowed to create a preliminary structure for the pool of collected institutions, thus generating a more systematic description and understanding of the inspected ecosystem of language actors.
In the second place, when deciding which documentation criteria were required to ensure a project-relevant depiction of the language actors, two factors were combined:

1. The intent of integrating the data into an online repository for user-friendly consultation, and
2. its potential for extension to also accommodate more and different institutions in the future.

[3] This exhaustive but concise listing (in terms of information for each institution) of local institutions is kept as part of the project documentation alongside the more detailed reports on institutions.

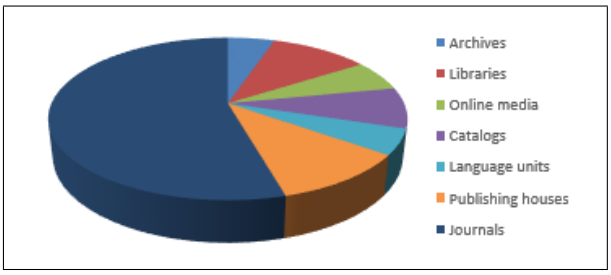

Figure 1: Distribution of institution types in preliminary classification

To this end, a paper questionnaire scrutinizing five main types of information about each institution was designed:[4]

1. background information on the institution
2. media collections
3. services
4. workflows
5. target groups

The reason why these key areas were considered as must-have descriptors is that they are the conceptual foundation and the structural pillars of the infrastructure: by analyzing and implementing them, recurrent patterns of interaction and interplay should emerge, be recognized and exploited to both establish and achieve synergies amongst the chosen organizations (Lyding et al., 2018).
This goal also guided the selection process for involving a small number of institutions and organizations into the project's information collection phase. While the setup of the documentation effort is designed to host an exhaustive description of South Tyrolean language actors in the mid to long term (see Section 6), the initial documentation effort described here aimed at describing a small sample of representative institutions to gain a first picture. Targeting an initial set of about ten institutions and organizations the following range of criteria for selecting the initial set of institutions were elaborated:

1. Quantitative and qualitative relevance in the territory,
2. category coverage – that is, at least one institution per type was selected with the aim of rendering an authentic, prototype-compatible cross-section of the chosen language ecosystem,
3. category descriptive completeness – for instance, for the type *library* three different organizations with non-overlapping media/data domains were selected[5] in order to present diverse facets of the category *library*, and
4. multi-category ascription – that is, institutions were selected which pertained to more than one category at once[6].

[4] See Appendix A for the listing of all aspects addressed.
[5] In this respect, a *general*, a *technical* and a *specialized* library were interviewed.
[6] A rather frequent combination involved institutions belonging to the categories "library" and "catalog" at the same time.

This allowed identifying cross-category features, thus understanding better the feasibility of describing complexity via a detailed documentation.
Out of the around 200 institutions recorded in the first identification phase (see above), eleven institutions have been selected, and were contacted one by one (see Section 4.2). All eleven institutions have been interviewed based on the structured questionnaire (see Appendix A): they served as a prototypical reference scenario for fine-tuning the preparation stage and, as a result, having a concrete database with a varied set of language institutions.

4.2. Establishing Contacts and Collecting Data

The data collection phase comprised three sub-stages: First the initial step of establishing contacts with institutions, second the step of explaining the project's objectives if the considered institution proved interested, and finally the concluding step of carrying out the interviews.
First, the selected organizations were approached to see whether they were willing and/or able to participate in the project and, if so, we involved them first-hand. Next, they were informed about the main objectives of DI-ÖSS and of the documentation process as well as of the procedure for collecting data. Also, possible questions from the side of the language actors were clarified in this step. From a communicative standpoint, we encountered a challenge in explaining the overall purpose of the project to organizations which have little direct experience in infrastructures. The explanations we provided moved from a goal-oriented overview of the system into its component parts, thus aiming at demonstrating the synergetic opportunities inherent to a language infrastructure project. From a practical viewpoint, describing how data needed to be acquired allowed the institution's contact person to make an informed decision as to whether or not to take part in the documentation initiative. It also helped the contact persons, to find their way of portraying the organization in light of the DI-ÖSS framework.
The actual data collection process was implemented through arranged interviews: they were conducted using the aforementioned questionnaire (see Section 4.1) as a code of practice. This guaranteed both content completeness – at least on a procedural level since some questionnaire areas had to be filled out flexibly according to each institution's specifics – and data collection standardization in view of creating facets (see Section 4.3). Furthermore, to maximize the potential of the meetings held with each selected institution, the following *modus operandi* was adopted: on the one side, a series of 'pre-investigations' were made into the organization by consulting its web pages; this permitted gaining initial contextualized impressions and, as a consequence, asking targeted questions during the interview. On the other, the encounter itself was audio recorded so as to collect data in as accurate a way as possible. Interviews were then formalized at a later stage. They were transcribed and during the transcription process new abstract attributes of description were identified (see Section 4.3). In particular, the transcription implied extensively completing the designed questionnaire in continuous prose and, where possible, adding links to existent institutional websites.

4.3. Presenting and Distributing Data

To present this inventory of language actors to potential users in the best possible way, the transcribed interviews were transformed into a more concise format resulting in a clear classification of the institutions. In this way a user can both easily identify the institutions they are looking for and, at the same time, explore the inventory for similar or related entries.

After careful consideration, it was decided to use a faceted interface. This should provide a good way of approaching the collected language actor data. Within CLARIN, the Virtual Language Observatory (VLO)[7], developed and maintained by CLARIN ERIC, provides the technical solution to a similar problem. It has to be said that the information being collected within the CLARIN VLO and DI-ÖSS slightly differs as far as content goes - the CLARIN VLO focuses on language resources, whereas DI-ÖSS looks at language institutions as a whole, including information on their institutional structure, resources (media collections) and services offered. However, the use case both projects are working on is still relatively similar in that it consists of collections of data which need to be presented in a compact and user-friendly way to make them accessible via the Internet. Apart from the VLO being well-maintained software and it being used in an important European infrastructure, this technical choice has the additional advantage of allowing for a future follow-up project in which the collection data - once separated from the general institution data - can be integrated into the CLARIN VLO. This will be much easier if the data in DI-ÖSS have already been collected in a CLARIN-compatible way.

Facet Name (German)	Facet Name (Translated)
Gattung	genre
Institutionsprofil	institution type
Rechtlicher Status	legal status
Digitalität	digital vs. analogue
Publikationszeitraum	publication period
Medientyp	media type
Automation	manual vs. automatized
Lokalität	local vs. remote
Dienst-Zielgruppe	service for target group
Sprache	language
Lizenz	licence
Workflow-Typ	workflow type
Software	software
Zielgruppe	target group

Table 1: Facets in the DI-ÖSS Language Actor Repository

To display the language actor documentation in the DI-ÖSS VLO, there first has to be a "translation" of the transcribed interviews into a more structured set of metadata and the most relevant metadata fields have to be identified and turned into facets that users can then use to filter the institutions.

Because of the exploratory nature of the project and the type of information that has been collected, the language actor documentation is very detailed and often shaped by the organization and work environment of the specific institution. However, some good candidates for facets did emerge from the data when we analyzed it specifically with this aim in mind. Overall we identified 14 relevant metadata fields with reasonable abstractions of their values, which we encoded into facets for the search (see Table 1). Examples of these facets are the 'institution type'[8], the sort of language data they mainly host (e.g. 'genre' *fiction* vs. *non-fiction*) or the time range of the items in a collection (i.e. 'publication period'). In addition a facet for filtering institutions by the services they offer was introduced (i.e. 'service for target group', such as *library catalog*).

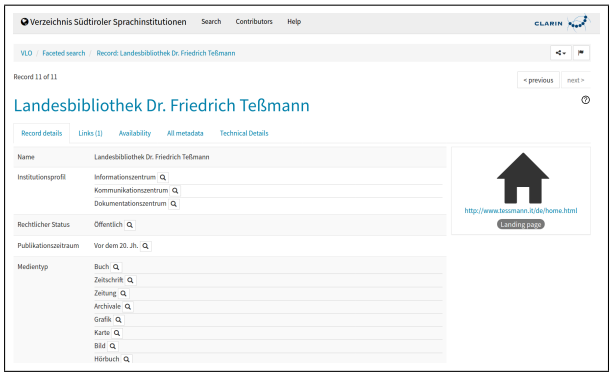

Figure 2: Repository entry for library Tessmann

Figure 2 gives the detailed view of the information recorded for the *Landesbibliothek Dr. Friedrich Tessmann*. Figure 3 shows the entrance page of the *Language Actor Repository* listing all institutions which are currently recorded.

Because of the severe information reduction that was necessary to transform the collected information into a format compatible with the VLO, it was decided that the full interviews should still remain available and be accessible so that users can always enquire more detailed information on an institution after they have narrowed down the selection through a faceted search.

On the more technical side, there had to be some preparatory work as well as editing of the VLO setup to make it work with our type of data. First, the metadata profile had to be formalized into a CMDI profile (Broeder et al., 2012) within the CLARIN Component Registry[9], so the VLO could process the data. As we consider our efforts not necessarily a part of CLARIN, we have decided to keep the profile that is being used in this project private for now. Then the VLO configuration had to be edited to support the facets selected for our project and at the same time superfluous standard VLO facets were removed.

The VLO software is provided in a Docker Compose setup[10] that could be installed without much additional work. It still needed to be slightly adapted for the use in this project. Apart from the facet configuration, the styling needed to be adjusted to reflect the project environment and make it more

[7] https://www.clarin.eu/content/virtual-language-observatory-vlo

[8] See Section 4.1 and Appendix A, questions related to background information about the institution.

[9] https://catalog.clarin.eu/ds/ComponentRegistry/

[10] https://gitlab.com/CLARIN-ERIC/compose_vlo

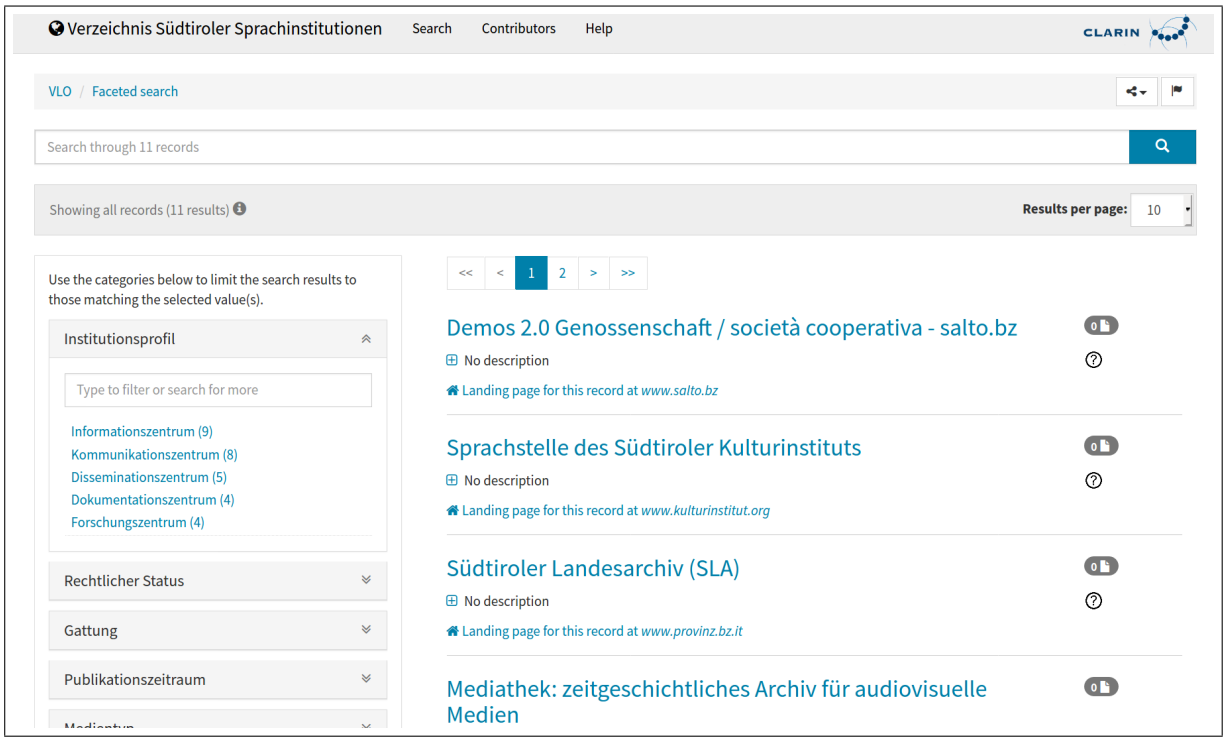

Figure 3: Entrance page of Language Actor Repository, with facet 'institution type' expanded

attractive for the envisioned target audience. Finally, there needed to be some technical adaptions to make the existing Docker Compose setup work on the Kubernetes cluster running at our institute. The DI-ÖSS *Language Actor Repository* is available online to the public at `https://kommul.eurac.edu/sprachinstitutionen/`.

5. Implicated Infrastructure Needs

Considering the experiences in this project and the reasons why it was set up, there emerges a vision for a larger-scale version of this kind of *language actor documentation*. Ideally this approach would be copied and transferred into other communities, documenting the actors in these environments in a similar way. As there is only one small pilot study so far we cannot be certain of possible unexpected obstacles, but if data collected elsewhere is comparable to the one found in the DI-ÖSS project it should be possible to aggregate this data on a higher level and CLARIN could set up a Language Actor Catalog as a companion to the Virtual Language Observatory. As described above there were certain difficulties with adapting the VLO software and especially the facets for the data collected about the language institutions. This suggests that the software might need some adaptions or a different, but similar, software should be used instead.

It is expected that setting up such a *CLARIN Language Actors Registry* (CLAR) will help smaller local institutions to more easily interconnect with other ones that are facing the same problems and could learn from each other in solving them. Also, it could help in finding institutions that face complementary problems, that means one institution has the solution to the other's problems and vice versa. Finally, having this repository at the European level means that these interactions and synergies can not only happen on a local level, but also across different countries. The same solution that works for a small historical library in South Tyrol might also work for a similar library in Catalonia, for example.

Additionally, by having this envisioned repository integrated within CLARIN, possibly using the CMDI standard for recording the data, it becomes much easier to take out just the information about the actual language data from the Language Actor profiles and integrating them into the CLARIN VLO. The VLO could then for each collection link back to the Language Actor Registry and in fact, this link could also be added for existing collections in the VLO (provided the institution has been added to the CLAR), where there is already a metadata field for this information called *Organisation*.

6. Summary and Future Work

In this paper, we have reported on a first attempt to create a comprehensive documentation of language actors in South Tyrol while raising awareness of the topic. In order to foster the growth and wide adoption of language research infrastructures, we claim that not only language resources and tools, but also actors in language-related domains need to be documented.

Within the DI-ÖSS initiative for South Tyrol, in this initial phase eleven institutions have been contacted and were interviewed in detail. The interviews were fully transcribed and information related to the selected facets of key information (see Section 4.3) were extracted and imported into the VLO.

In future work, we envision to extend the online documentation by populating it with information of the entire list of recorded language actors in South Tyrol (see Section 4.1) by asking them to fill short questionnaires related to only the key information encoded in the online documentation.

Recording the details of language actors allows both understanding their aims and needs and concretely mapping out

the language ecosystem on a general/global and local scope. To attain this goal, we suggest creating ways to grant access to data about language actors on a broader level and explore implications for the technical pre-conditions as discussed above.

7. Bibliographical References

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. (2012). Cmdi: a component metadata infrastructure. In Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme, volume 1.

Edmond, J., Fischer, F., Mertens, M., and Romary, L. (2017). The dariah eric: Redefining research infrastructure for the arts and humanities in the digital age. *ERCIM News*, (111).

Europeana Foundation. (2015). Transforming the world with culture: Next steps on increasing the use of digital cultural heritage in research, education, tourism and the creative industries. Technical report, Europeana Foundation, September.

Hinrichs, E. and Krauwer, S. (2014). The clarin research infrastructure: Resources and tools for ehumanities scholars. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Lyding, V., König, A., Gorgaini, E., Nicolas, L., and Pretti, M. (2018). Di-öss - building a digital infrastructure in south tyrol. In Inguna Skadiņa, editor, Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018. Linköping University Electronic Press. in press.

A. Questionnaire

The questionnaire collected the following types of information about South Tyrolean language institutions:

1. **Background information on the institution**
   (a) official name
   (b) year of foundation
   (c) address
   (d) official phone number(s)
   (e) official website(s)
   (f) library seal (ISIL)
   (g) number of employees
   (h) short description
   (i) historical background
   (j) institution type
   (k) legal status
   (l) organizational structure
   (m) contact person
   (n) main target group(s)
2. **Media collections** (separately for each collection)
   (a) type of collection: digital vs. analogue
   (b) data format
   (c) state of preservation
   (d) search database
   (e) genres covered
   (f) publication period
   (g) language
   (h) media type
   (i) copyright
3. **Services**
   (a) name of the service
   (b) short description
   (c) manual vs. automatized
   (d) local vs. remote
   (e) target group
4. **Workflows**
   (a) name of the workflow
   (b) short description
   (c) software used (including type, licence, short description)
5. **Target groups**
   (a) overall number of users
   (b) number of internal users
   (c) number of external users
   (d) short description (of each target group including their prototypical usage scenario
   (e) short description of secondary user groups
   (f) use case of main target group