DI-ÖSS - Building a digital infrastructure in South Tyrol

Verena Lyding and **Alexander König** and **Elisa Gorgaini**
and **Lionel Nicolas** and **Monica Pretti**
Institute for Applied Linguistics
Eurac Research, Bolzano, Italy
{firstname.lastname}@eurac.edu

Abstract

This paper presents the DI-ÖSS[1] project, a local digital infrastructure initiative for South Tyrol, which aims at connecting institutions and organizations working with language data. It aims to facilitate and increase data exchange, joint efforts in processing and exploiting data and synergies, and thus linking to big European infrastructure initiatives. However, while sharing the overall objectives to foster standardization, increase efficiency and sustainability, a local initiative faces a different set of challenges on the implementation level. It aims to involve institutions which are less familiar with the logic of infrastructure and have less experience and fewer resources to deal with technical matters in a systematic way. The paper will describe how DI-ÖSS addresses the needs for a digital language infrastructure on a local level, lay out the course of action, and depict the targeted short-, mid- and long-term outputs of the project.

1 Introduction

In recent years, the field of Digital Humanities has seen the development of multiple infrastructure projects at European level. Among the most well-known initiatives CLARIN (Krauwer and Hinrichs, 2014) and DARIAH (Edmond et al., 2017) target the needs of researchers, with CLARIN being mostly centered around the discipline of linguistics, and, to a lesser degree, history and literary studies, while DARIAH focuses on the broader field of all the arts and humanities. In some countries, like the Netherlands, CLARIN and DARIAH have even started to merge into a joint CLARIAH[2] project (Odijk, 2016).

Europeana, on the other hand, focuses on the cultural heritage sector (Europeana Foundation, 2015). Its main aim is to strengthen the networks between institutions like galleries, libraries, archives and museums (GLAM), especially by aggregating their metadata as much as possible to make them searchable in an easier and more convenient way. In doing this, Europeana also creates attractive portals to these data.[3]

The large field of smaller institutions, both in the public and the private sector, is not targeted by any of these big infrastructures, even though it could benefit from a close collaboration with Digital Humanities. It contains smaller libraries,[4] archives, cultural associations, and publishing houses; actors that deal with language and contribute to the field of research and heritage, but who are themselves too small to easily participate in one of the big infrastructures. These minor but central players are the target of DI-ÖSS: *Digitale Infrastruktur für das Ökosystem Südtiroler Sprachdaten und –dienste* (Digital infrastructure for the ecosystem of South Tyrolean language data and services).

2 Motivation

The increasing availability and the wide-spreading use of digital data in various academic disciplines, such as the humanities and the social sciences, have progressively led to heighten awareness to issues

[1] *Digitale Infrastruktur für das Ökosystem Südtiroler Sprachdaten und –dienste* - Digital infrastructure for the ecosystem of South Tyrolean language data and services

[2] https://www.clariah.nl

[3] For example, a special portal was launched to commemorate the start of the First World War. https://www.europeana.eu/portal/en/collections/world-war-I

[4] In contrast to the big national libraries targeted by Europeana.

Verena Lyding, Alexander König, Elisa Gorgaini, Lionel Nicolas and Monica Pretti 2019. DI-ÖSS - Building a digital infrastructure in South Tyrol. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 92–102.

about data standardization, preservation, exchange and reuse, thus calling for a shared agenda on workflows for data collection, processing and analysis. Consequently, a number of large-scale research infrastructure initiatives have been launched over the last decade (see Section 1). These have all been conceptually conceived as network communities and have primarily directed their efforts toward setting common standards, identifying best practices and employing technical solutions so as to foster accessibility, interoperability and sustainability whilst sharing and reusing data in national and/or international research contexts.

The project DI-ÖSS borrows potential from the aforementioned humanities infrastructures and functionally attempts to replicate their prospect of establishing connections and deploying synergies. Nonetheless, it theoretically adjusts such potentiality to a local level, namely to the Autonomous Province Bozen/Bolzano-South Tyrol situated in northern Italy. This shifts the operational focus with regard to both the types of participants/contributors and the scope/scale. Hence, small and very small institutions or companies, which are not necessarily connected to the research domain, turn into main actors, whereas the geographical, political and cultural area of South Tyrol becomes the core stage of the project.

The rationale behind this deliberate focusing and the subsequent course of action is twofold: in the first place, the pivotal role played by small organizations in performing fine-grained work on local cultural assets; in the second, the resulting need for notional models, actual practices and, as a linking element, tangible means of optimization in local contexts.

The first reason for upholding a locally-centered infrastructure is the contribution regional organizations make toward strengthening a sense of cultural identity and belonging. In fact, by systematically undertaking a series of data-driven tasks, i.e. collecting, recording, cataloging, processing, analyzing, evaluating, archiving and disseminating existing resources (cf. Figure 1) according to local needs, demands or requests, they buttress the conservation of today's ethnolinguistic legacy. However, given the broad spectrum the aforesaid duties span – from information retrieval to knowledge management – and given the independent recourse to consistent yet individually streamlined workflows, local actors' efforts often translate into heritage preservation and territory enhancement to a degree which may not be proportionate to their investments, and could be greatly facilitated by using synergies with other actors.

Moreover, regional data and services may exhibit particularities which are location-dependent and can, therefore, be effectively accommodated only at a local level.[5]

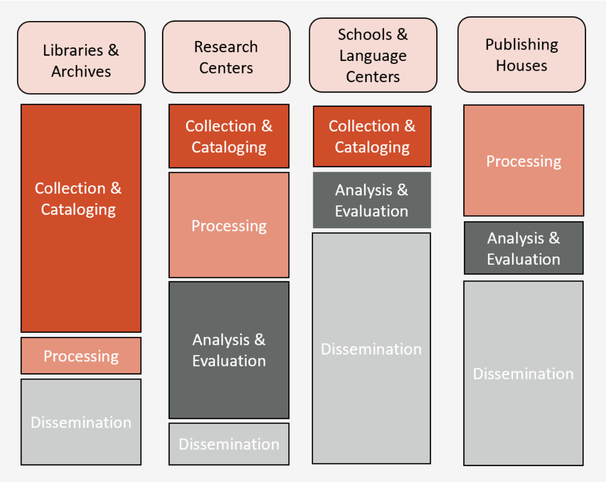

Figure 1: Exemplary overview of typical task allocation

The second motive for "going local" lies with the realization that regional institutions or companies are – as a tendency – prone to be less informed about existing approaches and/or ongoing efforts in the realm of infrastructure creation, development and implementation – which is currently pursued on higher structural levels (research, public sectors on cultural heritage preservation) and by larger frameworks, as

[5] E.g. South Tyrolean German is a local linguistic variety which is mostly documented and studied *in situ*.

explained above. They may, therefore, fail to benefit from the opportunities which arise in the process. Furthermore, should they succeed in keeping up to date, they might still lack some of the skills or resources essential to apply the knowledge acquired. By supporting small public or private organizations in the language sector in making the first steps to conform to bigger initiatives' standards, conventions and technologies, DI-ÖSS intends to sensitize them to the advantages of a digital language infrastructure. These include developing a framework for exchanging theoretical approaches and best practices, implementing specific interfaces for sharing data and/or tools, and coordinating and executing complex multi-step workflows. The primary objective is the promotion of cross-institutional efficient work.

In this regard, DI-ÖSS aims at actualizing the conditions for joining efforts, catalyzing processes and sharing outputs in order to support an interconnected ecosystem of South Tyrolean language data and services. Its synergetic potential is indicated by the existence of overlapping tasks and objectives amongst diversified organizations and by the use of comparable data sets. Having said that, a particularly apt way of exploiting such scope is a task-oriented allocation of work on the basis of each institution's dedicated spheres of competence, which allows for both medium-term quality improvement and long-term cost savings.

A concrete example may clarify this assertion. Libraries and archives specialize in collecting and cataloging text documents, whereas linguistic research institutes focus on analyzing data and developing sophisticated tools to automatically process and evaluate them. By combining and exchanging skills, the former can profit from rigorous high-quality linguistic research, while the latter can evade the elaborate task of data collection in return. Finally, data themselves prove valuable: by sharing them under established copyright rules, each institution can take advantage of a larger database without having to do any additional work to build it up.

3 Project Plan

As DI-ÖSS is a pilot project and deals with an abstract idea of a digital infrastructure, a careful project planning has been put into place to drive its step-by-step actualization and adaptation - if needed. Below this plan is laid out by first stating the aims of the project, then examining the approach to reach these goals and finally listing and illustrating the targeted outcomes.

3.1 Aim

The project aims to design and, in a second step, implement prototypical infrastructure components of a cross-institutional operational infrastructure, which is bound to digitally network South Tyrolean language data and services as integral elements of the local language ecosystem. This comprises different types of institutions and companies as well as four main project partners (see Section 4.1), all of which deal with language and cultural resources either in a commercial or in a non-commercial way.

It is within the consortium, selected to represent the wide-ranging variety of language institutions which the project sets out to cover, that DI-ÖSS plans to pilot the aforesaid prototype by means of concrete institution-specific use cases. They have been designed to meet the needs and reproduce the archetypal application scenarios of each partner, thus underpinning mutual bidirectional exchanges between organizations, permitting reusability in analogous circumstances or contexts and backing structural supportability in the project. Therefore, they ought to showcase how the target infrastructure can enrich and optimize each partner's contribution.

Additionally, the implementation of the infrastructure itself and of the correlated use cases is geared toward time-staggered objectives. In particular, the short-term facilitation of outputs at a local level should augment the quality of the project consortium's work, generate room for further development and lead to the medium-term extension of the infrastructure to the entire South Tyrolean language ecosystem. Then, the long-term evaluation of the infrastructure feasibility will be accompanied by the overarching aim of connecting it to national and pan-European initiatives.

Lastly, the higher-level goal of DI-ÖSS is, among other things, laying the foundations of the South Tyrolean digital cultural heritage, as the project motto epitomizes: "Start local, think big."

3.2 Approach

The aforesaid large-scale projects – CLARIN[6] at the forefront – are transnational initiatives of crucial reference for DI-ÖSS in that they offer efficacious examples of location-independent aggregation, convention-driven harmonization and user-friendly utilization of valuable resources. They, therefore, provide a yardstick by which to draft preliminary aims and orientate expected results. This liaison notwithstanding, DI-ÖSS differs from them in a series of aspects,[7] the foremost being the theoretical approach (for an illustrative exemplification of how CLARIN and DI-ÖSS differ in their specific focus cf. Figure 2).

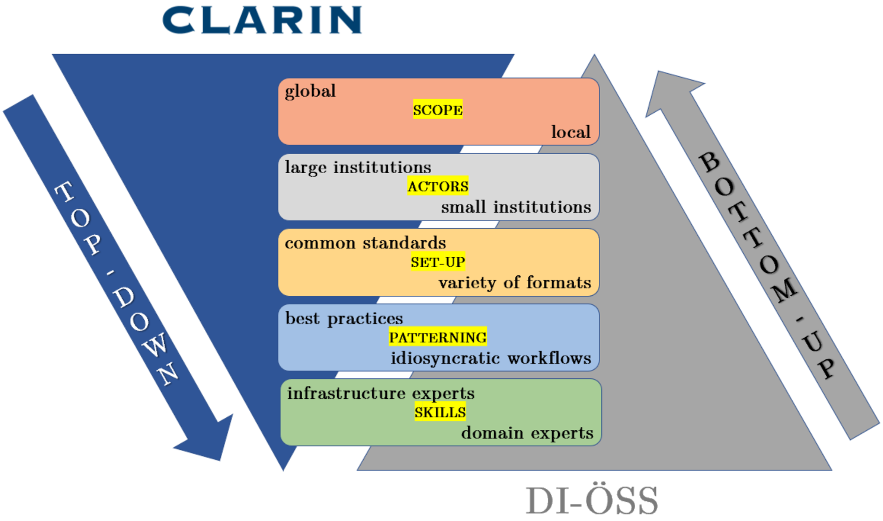

Figure 2: CLARIN vs. DI-ÖSS

Bigger projects employ a top-down method in accordance with each initiative's agenda, developing intelligible research applications. Conversely, DI-ÖSS makes use of a bottom-up strategy, intentionally launched by addressing a more-applied type of participants and a local scope. This anticipates two further points: on the one side, the necessity of *following the actors* (Latour, 2005), i.e. *learn from them* so as to identify data and service overlaps and share solutions by means of functional interplay and technical interoperability. On the other, the challenge of incorporating a heterogeneous ensemble of contributors – consisting of diverse data formats, idiosyncratic workflows and domain experts – into a coherent *assemblage* (Deleuze et al., 1987), i.e. a multiplicity of configurations connected in fluid exchanges. DI-ÖSS intends to pursue the surfacing of plastic articulations amongst the participating institutions and companies so that relevant theoretical and practical resources can flow within the infrastructure and give rise to a dynamic *constellation* of placeable, displaceable and replaceable relationships of *exteriority*.

3.3 Expected Results

As far as the project is concerned, we are targeting the outputs listed below with the intention of finding a balance between workable expectations and practical opportunities:

- Establishing a competence-based purpose-built cooperation which assembles and maintains language data and services;
- Adopting shared procedures and standardized data formats;
- Connecting institutions, companies, their data, services and search portals in integrated fashion and in view of relevant copyright and economic requirements;

[6] As mentioned in Section 1, CLARIN is but one of such projects. Nonetheless, given its exemplarity, it is here regarded as a touchstone to compare and contrast DI-ÖSS with well-established European infrastructures.

[7] The differences in terms of project scope and actors' participation have already been highlighted in Section 2.

- Providing a stable, centralized, flexibly expandable work environment;
- Restructuring redundant work steps and consequently reorganizing personnel responsibilities so as to recuperate misplaced resources and reinvest them.

The coalescence of these potentialities into an interoperable network is anticipated to lead to a collective advancement in terms of innovation. In fact, from a general perspective, DI-ÖSS ought to generate added value in form of increased and improved quality, efficiency, visibility and sustainability of the data and services offered. The overall character of these advantages outlines their versatility and malleability, i.e. leeway for the stakeholders to resourcefully mold or shape them into personalized benefits according to their institutional interests, needs and aims (for an exemplary overview of added value adaptation cf. Table 1).

Type of Stakeholder	**Added Value**
Libraries and archives	Use of computer linguistic tools and data processing stages
Research centers	Access to carefully documented collections of texts
Schools	Access to search portals with enhanced functionality
Language centres and cultural institutes	Access to carefully documented collections of texts and use of data analysis tools
Institutions and organizations responsible for education policy	Access to extensive studies and observations on the language situation in South Tyrol
Publishing houses	Use of automatic data analysis tools

Table 1: Overview of added value adaptation

Finally, the results of every project phase will be compiled into an evaluation report and a plan for a follow-up project aimed at building a comprehensive and sustainable digital language infrastructure.

4 Implementation of the project plan

The project, which was set in motion on January 1st 2017 and has been running since, is organized along a number of phases which build one upon the other. In the beginning, a small project consortium is built, which includes institutions from each of the most relevant target groups within the local language ecosystem (see Section 4.1). At the same time, detailed information regarding language data and services is collected from a wider set of institutions and organizations in South Tyrol (see Section 4.2). Afterwards, building on the previous phases' insights, specific use cases for each partner institution are determined and defined in greater depth. Along these use cases, a set of prototypical infrastructure components is built by implementing the technical setup required to connect partners and their data as needed by each use case. Finally, the infrastructure is piloted by employing it for each of the use cases and improving technical and conceptual aspects within the process (see Section 4.3).

4.1 Consortium

The project consortium was built during the first months of the project. In that initial phase the project partners were chosen carefully so that a range of different types of institutions as wide as possible could be represented in the consortium. As the DI-ÖSS project is aware of and embraces the diversity of relevant language institutions in South Tyrol the various relevant institution types had to be taken into account when considering potential partners for this pilot project and its envisioned infrastructure. The selection process has therefore been looking at a number of different institutions and companies, especially considering their approach to the development, distribution and preservation of language data and services. As explained in more detail later (see Section 5.2), due to its status as a pilot, the project can be perceived as abstract, thus making it challenging to communicate its objective and potential to the targeted partner institutions.

The final project consortium has been established between the following four institutions:

1. the *Institute for Applied Linguistics at Eurac Research* (project lead) as a research institution working with empirical language data and related language technologies,

2. the *Landesbibliothek Dr. Friedrich Teßmann* as a general-purpose library with a large digital collection of texts,

3. the *Sprachstelle* (”language unit”) of the South Tyrolean Institute of Culture as a central institution for promoting the German variety of South Tyrol and informing the public about related matters,

4. the news and community portal *salto.bz* as a South Tyrolean publisher of daily news, local content and discussions around it.

4.2 Stocktaking - *Bestandsaufnahme*

The DI-ÖSS project, started of with the “Bestandsaufnahme” (literally *stocktaking*), which designates an initial work phase with a distinctively informational character. It involves collecting and thoroughly categorizing facts and details on the project partners and participating organizations with the intention of mapping the current state and gaining some insight into the nature of their data and *modus operandi*.

With this end in view, a questionnaire concerning five key aspects, i.e. general information on the institution or company in hand, its data collections, services, workflows and target groups, has been developed so as to cover the major fields of interest for the end users.

1. General information: on the one hand, it provides a global overview of the type of organization selected; on the other, it describes its specific needs and wants/intents and purposes while allowing each institution to become part of the infrastructure itself.

2. Collections: it presents the analogue (print) and/or digital data sets typical of each organization and expands upon them by differentiating between content and technical data. The former include criteria, such as a genre-based sorting of the material (principally fiction vs non-fiction), its amount and language of composition; the latter comprise parameters, like a medium-based sorting of the material (e.g. books, newspapers, journals, etc.), its format(s) and the software(s) used internally for working with it in the broadest sense. Moreover, this section contains indications as to which copyright terms and conditions apply for each collection.

3. Services: it labels the main, user-tuned, institution-specific services offered and briefly describes them, i.e. interlibrary loan, archive research inquiries, etc.

4. Workflows: where applicable, it sketches internal procedures concerned with data management, i.e. acquiring, processing and disseminating them, and service provision.

5. Target groups: it categorizes principal and secondary user groups, their approximate size and, if possible, how these typically make use of an institution’s data and/or services.

The information collection process has so far taken place in the form of a recorded interview whose content is minuted at a later stage. In a second step, the key information is copied into a CMDI XML[8] document that is based on a profile adapted for the project’s specific needs and is then fed into a modified VLO[9] where the data can be browsed via facets. These procedural steps have been carried out factoring in both the plausible future integration of the specifics into a larger CLARIN-like language resource infrastructure and especially the possibility to ingest parts of the data into the actual CLARIN VLO so as to make it more visible to the larger research community.

[8] https://www.clarin.eu/cmdi

[9] http://www.clarin.eu/vlo/

4.3 Use Cases

Specific use cases are identified for each project partner. The use cases are selected and defined in order to best comply with the following four aims:

1. Enhancing a task in the partner's daily workflow;
2. Exploiting a synergy (shared or complementary expertise) with at least one other partner;
3. Being applicable or easily adaptable for future or similar tasks;
4. Allowing to build a generic infrastructure interface for handling them

In the following subsections we will briefly depict the four use cases.

4.3.1 Use Case 1 - Teßmann library: browsing cultural magazines

Use Case 1 addresses the task of serving *enhanced search facilities for digitized content* to library users. It is built on a collection of cultural magazines from the 70s and 80s which contain written content and pictures, and follow an irregular layout structure (e.g. paragraphs of texts are blended with images, articles run over several – not always consecutive – pages, etc.). Content-wise the magazines combine cultural reviews, lyrical and poetic contributions, portraits of artists and artworks as well as announcements, manifests and reports. Accordingly, readers would prefer to browse them based on recurrent themes, figures or also concepts. For example, a library user might want to find all articles related to one artist or a prevalent topic of discussion. In addition, locations, time periods or arts genres might be themes of particular interest.

Approach

The delivery of enhanced search facilities is approached in three steps. First, the digitized texts are processed and annotated for information of interest, such as thematic keywords, persons, locations, arts genres, etc. Second, the individual articles of the cultural magazines as well as smaller text snippets are interlinked, using the annotated information (step 1). Third, a search interface that allows browsing of the cultural magazines based on the presented concepts and their interlinking is created. For example, the interface offers the user access to related articles grouped together and navigation along related concepts or interconnected persons and themes.

Synergies

In order to implement the use case, computational linguists at Eurac Research are closely collaborating with experts in literature and cultural studies at the Teßmann library. The literary and culture study experts work on identifying themes and aspects of relevance and clearly describe their informational needs directed toward the cultural magazines. Based on these pointers, the computational linguists select and apply NLP tools to automatically detect and annotate these types of information in the texts. Finally, in close collaboration they design and implement an interactive interface for searching the texts based on the advanced textual cues.

The use case is non-specific and transferable to the extent that a generic toolchain for annotating digital texts is put into place, and a search interface that builds on the annotated text formats is created so as to access and display segments of text (magazine articles or smaller paragraphs).

4.3.2 Use Case 2 - salto.bz: enhanced tagging of articles

Use Case 2 is concerned with the task of *improving search and discoverability* for the readers of the online news portal *salto.bz*.

Currently, the portal offers readers the built-in search of their CMS, which only performs simple string matching and no ranking of the search results. This makes finding interesting content much more difficult for the readership, both targeted search and browsing by being offered similar articles are not very efficient at the moment. Most readers will likely browse through the various sections without looking for something specific, they could therefore benefit from articles being more closely interlinked so that related articles could be offered automatically, giving more background or a different view on a subject. But also the targeted search is a valid use case for readers of a news site. Ideally, it would be possible to limit search results to a specific news section (politics, sports, local news, etc.) and also to a specific time-frame.

Approach

The problem is approached in a number of separate, but closely interrelated steps. All the existing articles will undergo a semantic analysis to extract the most relevant keywords and the same will be integrated into the editorial user interface so that the mechanism can also be triggered for newly written articles. At the same time, the semantic analysis will automatically identify related articles and link them directly to each other. This newly generated deeper information about the content of the articles will then enable a much more user-friendly search interface to be implemented. One especially challenging part of this endeavor is the fact that the news portal in question is bilingual in scope. Both articles in German and Italian are being published and a well-designed search has to take into account that the readers would like to find results in both of the languages no matter which language they are using to search. This means whether someone is searching for *elezioni* or *Wahlen* they should obtain the same set of search results.

Synergies

The implementation is taking place at two different ends of the use case. While computational linguists are working on the backend and creating web services that are able to automatically extract keywords and compare articles for relatedness, the editors and technicians of salto.bz will deal with the frontend. The editors are checking and approving the automatically generated keywords and relations between the articles. At the same time, the salto.bz developers will adapt the user interface to make the newly generated keywords accessible to the readers and create an improved search interface that makes use of the additional information on the articles that is now available. As the web services that offer the linguistic services will be implemented using a generic API, they could in principle be used by other interested parties at a later stage.

4.3.3 Use Case 3 - Eurac Research: crowdsourcing of historical letters

Use Case 3 is a cooperation between two institutes of Eurac Research, the Institute for Applied Linguistics (IAL) and the Institute for Minority Rights (IMR). The IMR has been collecting missives (mostly letters and postcards) from the inhabitants of South Tyrol to create a representative corpus of historical letters spanning the 20th century. Within this use case, the still-growing collection of mostly handwritten letters and postcards will be enriched with structured metadata and transcribed by the local population using online crowdsourcing tools. During the crowdsourcing phase, the public will also be involved through public events and the envisioned end result is both a well-curated digital collection and a highly engaged and passionate (part of the) public.

Approach

There is already a huge collection of some twelve thousand missives that have been digitized as pictures, while more material is still being collected. In a first step, the data will be uploaded into an instance of the crowdsourcing software Pybossa where volunteers can extract the metadata (sender, addressee, date, etc.) from each item which is then stored in a structured machine-readable format. After the metadata have been extracted, the missives can be grouped into related collections (e.g. based on location or time) and be ingested into a web-based annotation software. The project is using the web version of Transkribus[10] to crowdsource the transcription.

Synergies

The Institute for Minority Rights is collecting the missives from citizens all across South Tyrol and takes a first step of digitizing them. The whole technical setup of the various crowdsourcing tools is handled by experts at the Institute for Applied Linguistics while the design of the user interface and

[10] https://transkribus.eu/r/read/projects/

the accompanying texts that guide the users are jointly developed by both institutes. This collection of historical letters is a great opportunity for such a shared project because the content, while being very interesting from a historical perspective on the eventful 20th century in South Tyrol, also offers an insight into a unique linguistic situation where writers often switch between standard German and their local dialects, sometimes in the same letter. And especially after the annexion of the territory by Italy the language mix that can be found in the texts also includes Italian. Additionally, during the war periods we can find a lot of letters by authors that are not very used to writing longer texts, which also promises to yield some interesting analyses.

With the IMR's focus on making this part of their cultural heritage available to the population of South Tyrol, all data (as far as the obvious privacy concerns allow) will be made freely available on the internet so that also hobby scholars and citizen scientists can use it for their own studies. It is also expected that the experience both institutes will gain in the area of crowdsourcing will be beneficial to future projects.

4.3.4 Use Case 4 - *Sprachstelle*: identifying regional neologisms

The final Use Case is a project aiming at installing an infrastructure for finding and identifying neologisms that are specific to the region of South Tyrol. For this use case, computational linguists automatically harvest South Tyrolean sources on the internet to propose candidates for such neologisms and experts on the local variety of German at the language unit (*Sprachstelle*) of the South Tyrolean Institute of Culture will verify which of those candidates can actually be seen as potential regional neologisms. This feedback will then be used to fine-tune the automatic detection to minimize the amount of manual work that has to be done by language experts further on.

Approach

The starting point for the use case is a carefully curated list of South Tyrolean media that publish original texts online. Among those are web sites of newspapers and local TV and radio stations, but also personal or semi-professional websites or sites that provide information from the local government. This list of websites is then regularly crawled and the resulting word list is checked against a list of standard German words to eliminate known forms. After trying to automatically eliminate as much as possible also errors resulting from the known error-prone process of crawling HTML pages, the list of candidates for possible neologisms is then checked by experts on the local variety of German to determine if a candidate is a neologism and if so, if it is specific to the linguistic variety spoken in South Tyrol. The edited list is then fed back into the algorithm that selects the candidates from the web crawl resulting in a continually improving selection process.

Synergies

The Institute for Applied Linguistics at Eurac Research has implemented a first version of the software that crawls the web and selects possible neologism candidates, called *Styrlogism*. First editing rounds have already carried out internally with experts on South Tyrolean German from the institute itself. Within this use case, this will now be complemented by the expertise coming from the language unit at the South Tyrolean Institute of Culture. As the main task of the language unit is to inform and educate the population about the local variety of German, they can use the results from this process to showcase the newly detected South Tyrolean neologisms in their public relations work. In this presentation, it is often possible to show the original context of this discovery because the Styrlogism tool keeps the whole environment of the detected words and also always stores the originating website. If this has not been taken down by their authors in the meantime, interested users can then even go back to the original source to see the new word in the complete context.

5 Discussion of encountered challenges

Even though conceptually aligned with established large-scale infrastructure initiatives like CLARIN, the actualization of the locally-oriented DI-ÖSS language infrastructure is a step into uncharted territory. Especially the fact that the language partners involved and approached within the course of the project are relatively small, have few resources at their disposal and possess limited experience with large-scale projects has proven to pose particular challenges, which could not be anticipated to the extent encountered. The DI-ÖSS project is devised as a pilot project that is specifically designed to find the unique challenges inherent in such a local infrastructure. The goal is to learn from the prototypical phase and use it as a facilitator to establish a comprehensive and powerful digital language infrastructure in South Tyrol in the mid to long term and take steps to integrate it as much as possible with larger infrastructures like CLARIN and DARIAH.

Having said that, we will close this article with a discussion of some of the most prominent challenges that we have faced over the course of the project up to now. They concern conceptual, communicative and technical aspects, as laid out in the following three subsections.

5.1 Conceptual challenges

Already when creating the consortium, but especially later when interviewing potentially interesting institutions for the *Bestandsaufnahme*, it became apparent that while everything can be considered potentially interesting linguistic data – from the protocols of the province offices to advertisements of a local company – DI-ÖSS has to tighten its scope to more obvious ”language institutions” like libraries, publishing houses and linguistic research centers (see Section 4.1). Generally, the focus was reduced to institutions that 1) deal with language data produced in South Tyrol, 2) consider working with language data their main activity, and 3) work with data that are available digitally, either digitized or born digital. It was also decided to explicitly involve smaller actors that do not already have visibility and power in the South Tyrolean ecosystem in order to make the resulting infrastructure more of a democratic place. Additionally, there was a problem with clearly defining some of the use cases. On the one hand, a very deep understanding of the workings of an institution is fundamental to see whether there are specific needs; on the other, a wide knowledge of the other partners is essential to determine which possible solutions there could be to those problems. This was only solvable by taking the time to have long discussions with each project partner to fully understand their needs and capabilities.

5.2 Communicative challenges

Communicative challenges arose in the process of getting institutions interested in joining the project as it has proven difficult to properly communicate the scope and purpose of it. As described in Section 3.2, the infrastructure can be theoretically seen as a fluid assemblage of actors, which influences the South Tyrolean culture and identity, and sets the basis of a South Tyrolean digital cultural heritage. Translating these theories into graspable concepts for the possible partners is certainly a challenge. It helps to use metaphors of physical infrastructures, like the railway system, and also to focus on concrete use cases early on, so that it becomes easier for the potential partners to see their specific role within the project.

It is necessary to address every possible partner institution with a different approach, trying to anticipate their needs and reservations. While libraries and other public institutions are more readily willing to share their data freely, commercial actors, e.g. publishing houses, are often very protective of their data as this is central to their business model. But even once a potential partner sees the benefits of the project, there are still further issues. Because of the small size of many of the potential partners, there might not be enough resources available that they could bring into the project. Especially, if there is no obvious short-term benefit for the partners, it becomes difficult to justify spending some amount of their often quite limited human resources on this project.

5.3 Technical challenges

This is another point where this small-scale infrastructure differs considerably from its larger counterparts. Many language partners in DI-ÖSS have very limited resources, both on the personnel and on the IT side, so it usually is difficult for them to implement large changes in their data management infrastructure or their typical workflows, while this is more feasible for the bigger institutions involved in infrastructure projects like CLARIN. This means the DI-ÖSS infrastructure has to be constructed in such a way that it integrates the needs of an infrastructure (standardized data formats and APIs) with the existing working realities, which often involve suboptimal or home-grown solutions that cannot be easily changed or adapted.

References

Gilles Deleuze, PF Guattari, and Felix Guattari. 1987. *A thousand plateaus: Capitalism and schizophrenia*, volume 19. University of Minnesota Press.

Jennifer Edmond, Frank Fischer, Michael Mertens, and Laurent Romary. 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.

Steven Krauwer and Erhard Hinrichs. 2014. The clarin research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531. European Language Resources Association (ELRA).

Bruno Latour. 2005. *Reassembling the social: An introduction to actor-network-theory*. Oxford university press.

Jan Odijk. 2016. Clariah in the netherlands. In *LREC*.