Proceedings of LR4SSHOC: Workshop about Language Resources for the SSH Cloud, pages 33–36
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

LR4SSHOC: The Future of Language Resources in the Context of the Social Sciences and Humanities Open Cloud

Daan Broeder[1], **Maria Eskevich**[1], **Monica Monachini**[2]
[1] CLARIN ERIC, [2] Institute of Computational Linguistics - CNR
[1] Utrecht, The Netherlands, [2] Via Moruzzi,1 – Pisa
d.g.broeder@uu.nl, maria@clarin.eu, monica.monachini@ilc.cnr.it

Abstract

This paper outlines the future of language resources and identifies their potential contribution for creating and sustaining the social sciences and humanities (SSH) component of the European Open Science Cloud (EOSC).

Keywords: Language Resources, European Open Science Cloud, Social Sciences and Humanities

1. Introduction

The term Language Resource (LR) refers to a broad type of speech and language data in machine readable form, used to study language or assist language processing applications. Examples of Language Resources are: written or spoken corpora and lexica, multi-modal resources, grammars, terminology or domain specific databases and dictionaries, ontologies, multimedia databases, etc. Language Technology (LT) are broadly defined as software tools for the analysis and use of Language.

The development of LR, LT, and their management have reached the level of maturity that allows their application to be expanded beyond the borders of traditional linguistic disciplines. In principle, such proliferation of resources and technologies is at the core of initiatives leading to the creation and building of the European Open Science Cloud (EOSC)[1]. EOSC is the latest development of the European approach to research infrastructure building. It has a long history starting with the series of ESFRI roadmaps[2].

In every field of research the steps towards EOSC have led to discussions on how to best accommodate the needs and requests of representative communities while complying with the technical requirements inherent to the implementation of EOSC.

At previous stages the ERICs (European Research Infrastructure Consortium)[3] were established to answer the need of specific research communities for infrastructure solutions. The first two to be created were SHARE[4] and CLARIN[5], both ERICs for the social sciences and humanities (SSH) domain. This early uptake illustrates the overall involvement of the SSH in shaping the European Research Infrastructure landscape.

Nowadays the SSH domain has grown a number of research infrastructures (RIs), CESSDA[6], CLARIN, DARIAH[7], ESS[8], SHARE[9], that support their domain-specific work with examples of collaboration where linguistic analysis is used to support studies into societal and cultural dynamics. Research in the broader SSH domain can benefit from language data because of the potential value of extracting information from data expressed in the form of natural language. Well organized and easy access to language resources and relevant processing tools can stimulate the broadening of research questions and the improvement of tools for the analysis of language resources.

Some SSH research infrastructures have been able to articulate their community requirements and their domain-specific agenda by participating in large e-Infrastructure projects, such as EUDAT[10], EGI[11] and EOSC-hub[12]. The latter project has integrated some key CLARIN services[13] into European Open Science Cloud (EOSC).

To foster collaboration between related research domains, the European Commission (EC) has introduced funding instruments for thematic cluster projects which are open for consortia consisting of multiple ERICs from related disciplinary fields. Cluster projects are expected to develop common solutions for similar problems and inform one-another about specific approaches.

The project Social Sciences and Humanities Open Cloud (SSHOC[14]) is such a cluster project, which is part of the series of INFRA-EOSC projects: H2020 initiatives aimed at building the EOSC, and the to creation and support of the SSH part of the EOSC via alignment and integration of infrastructural services from the Social Sciences, Humanities and Cultural Heritage. SSHOC represents the third generation of SSH cluster projects.

This paper describes the role of LR and LT in the context

[1] https://ec.europa.eu/info/publications/european-open-science-cloud-eosc-strategic-implementation-plan_en
[2] https://www.esfri.eu/esfri-roadmap
[3] For an overview of all ERICs established and the links to their websites, see the information pages of ERIC Forum on the ERIC Landscape.
[4] http://www.share-project.org
[5] https://www.clarin.eu
[6] https://www.cessda.eu
[7] https://www.dariah.eu
[8] www.europeansocialsurvey.org
[9] http://www.share-project.org
[10] https://eudat.eu
[11] https://www.egi.eu
[12] https://www.eosc-hub.eu
[13] https://www.clarin.eu/eosc
[14] https://sshopencloud.eu

of the SSH Open Cloud. Section 2 addresses the main features that define the development of research infrastructures at large are discussed. Section 3 zooms in into the specific aspects of European long-term existing initiatives that build the ground for future development. Section 4 identifies the current capacities and current difficulties in sharing and optimising research data and creating and sustaining an infrastructure for the SSH domain, and Section 5 summarises the most prominent issues and potential that will define the configuration of LR and LT in the context of EOSC.

2. Challenges of RI landscape for the SSH

As outlined in Section 1, the EC initiated EOSC that is currently rolled out in the form of a number of European Union (EU) level and regional projects. In order to understand the dynamics of such developments, it is useful to examine diverse challenges in the current research environment that influence the decision taking process, and that are expected to be tackled through EOSC. The very dynamic landscape of research infrastructure builders and service providers is influenced by a number of trends and interests:

- Data deluge: Although currently already a well described, "almost flogged to death" concept, it is still a challenging and unsolved in daily research reality, and requires the uptake of (for many) new highly performant data management solutions. (Hey and Trefethen, 2003)
- Scale up for efficiency: On the one hand, the goal to serve targeted research communities is achieved through creation and development of ERICs and thematic cluster collaborations that are expected to provide more generic services. On the other hand, the non-thematic service providers, e-Infrastructures as EGI and EUDAT that do not serve a particular community understandably tend to avoid diversification of their services.
- Professionalising service provisioning and software development: There is a trend to follow technology adoption and operational protocols from industry, as for instance the use of IT Service Management as FitSM[15] and the use of security standards as ISO 27001[16], which is in agreement with the previous point. In general this can be a beneficial development improving efficiency.
- **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (Wilkinson et al., 2016): Across different research domains there is a broad agreement to adopt FAIR principles when implementing services and data management protocols as a means to ensure proper access and re-use of research data and services.

In addition to the above listed trends that can be observed across all domains, there is number of domain- and community- specific aspects of research infrastructure building that have intrinsic influence on the decisions taken:

- While there is a steady increase in technical knowledge and proficiency amongst the researchers in the field of SSH, it appears that a number of intrinsic limitations in terms of technical complexity of research software and infrastructure solutions are still in place. The SSH community as a whole does not accept high IT proficiency as a default requirement for engaging into research. For instance purely qualitative researchers usually do not make use of advanced technical solutions. In comparison to "hard sciences", it is harder to engage with and to encourage the SSH community as a whole to change the research methodologies adopting broader usage of technological solutions, and there is an urgency to make the infrastructure developers aware of the need to communicate with end-users at the appropriate level of expertise. Note that scaling-up IT support organisations will not always favour easy support and communication with end-users.
- Another aspect of current SSH infrastructures that is important to consider is the use of strong thematic centres as a backbone. This is the case for at least two RIs: CESSDA and CLARIN. These centres play an essential role within the research infrastructures, hosting data and services for their own purposes and users, but also, as in the case of for example CLARIN, contributing to the central infrastructure services. The thematic centres are largely funded by individual national governments and funding agencies and are major national contributions to the a common European infrastructure. The RIs play an important role in formulating quality and certification requirements for those centres, in accordance with EOSC policies, that they themselves helped to shape.

3. SSHOC contribution and impact

Realising the SSH part of the EOSC, means, on the one side, to make SSH data, language processing tools, and services available, adjusted and accessible for users across SSH domain and, on the other side, to align and integrate services and infrastructures from the Social Sciences, Humanities and Cultural Heritage with one another and with the now emerging EOSC.
Different perspectives on activities, contributions and results aimed at realising the SSH Open Cloud can be outlined:

- From the perspective of EC: it is important to know that the thematic cluster projects such as SSHOC were specifically intended to make a link between the EOSC development and the research communities and that cluster project play an important role in the community consultation process and EOSC governance.
- From the perspective of a thematic cluster project such as SSHOC: challenges and successes should ideally be measured by how project results can be made useful for the common good, i.e. the outcome should be useful for more than just one of the SSHOC stakeholders.

[15] https://www.fitsm.eu
[16] https://www.iso.org/isoiec-27001-information-security.html

- From the perspective of a SSHOC partner: the possibility for the individual stakeholder infrastructures to build long term partnerships and to make related initiatives known and accessible to the other SSHOC partners is important.
- From the perspective of communities that are served by RIs within the thematic cluster: there is a possibility to be represented and to have the community voice/opinion heard at the level of EOSC through SSHOC activities. For example, the larger SSH communities, such as DARIAH, E-RIHS[17], CESSDA, can benefit from CLARIN's long-term relation with the European and US Language Resource agencies, such as ELDA[18] and LDC[19] (Cieri, 2020), and CLARIN involvement in LREC community. Whereas the LREC community and representative agencies expose their resources and tools to large audiences of potential users.

As SSHOC can offer a link to and a view on how the EOSC initiative searches to transform the way research infrastructures are built and made available to researchers. This can have consequences for the way Language Resources and Technology (LRT) should be produced and made available. Thus, it can be considered as part of the CLARIN mission in SSHOC to consider and discuss such consequences for the traditional LRT centers as it should for the CLARIN centers.

While national organisations involved in CLARIN consortia and SSHOC are also participating in other European initiatives with a focus on LT, such as the industry-oriented ELG project[20], their participation in SSHOC can also contribute to a better alignment with EOSC. Thus it helps EOSC to generate further impact outside academia.

One of the benefits for the infrastructures participating in SSHOC is the sharing of infrastructure building efforts amongst partners and the potential for developing a common strategy towards the landscape dynamics and trends listed in Section 2. The possibilities for scaling up the use of infrastructure components through SSHOC is illustrated by the planned uptake of two key CLARIN infrastructure components by DARIAH, CESSDA and E-RIHS, in this case the CLARIN Language Resource Switchboard[21] and the CLARIN Virtual Collection Registry[22]. Extensive consultation between SSHOC partners takes place to guide their generalisation and find integration opportunities. Another example is that after an evaluation of the vocabulary management platforms currently used in the SSH domain and the selection of one or more common registries and a common management platform, wider visibility of agreed vocabularies are likely to be expected.

Also with respect to the creation of completely new common infrastructure components, there are two planned examples: (i) the SSH Open Marketplace mentioned below and (ii) a prototype of a SSHOC Citation infrastructure for "FAIR SSH Citations" which is intended to make citations machine-actionable.

4. Aspects of LR4SSHOC implementation

In order to bring the LR to the SSH Open Cloud diverse aspects of implementation are to be taken into account.

4.1. Findability of services and solutions

Although the need for stable well defined data management and processing services for research cannot be denied, equally or even more important is the need for flexibility and short turn-around with regard to implementation of new requirements and adaptation of for instance natural language processing (NLP) software and workflows to new insights emerging in scholarly practises or from thematic research agendas. In such cases the current solutions for service registration and evaluation offered by the EOSC solutions, the EOSC catalogue and EOSC Marketplace are probably too inflexible for the integration of domain-specific services, and clearly more suited for generic data management services that are typically very stable. However, registration of services is crucial for wider visibility, and therefore, it is important to enable registration for the often more dynamically developing class of domain-specific research software. Not only for sharing amongst researchers but also for acknowledgement and visibility by funding agencies. However, service registration would be more effective if the agency is closer to the research communities that may be also better positioned to contextualise the registered services in terms of solutions for specific research problems. SSHOC is developing the SSH Open Marketplace also to address this need of explicit registered services in a community-managed fashion. Obviously, such thematic service and solution registries will require proper funding and especially sufficient editorial support by the communities.

4.2. Interoperability

Interoperability of data and services is the holy grail in many research infrastructure plans. Within the SSH, the interoperability with respect to data formats is probably the more easy goal to achieve, as the RIs focusing on LT and LR in the SSH context (CLARIN and DARIAH) already have adopted international standards for the use of important data formats and shape jointly the common standards, such as Text Encoding Initiative (TEI)[23] and other annotation formats. However, with respect to solving semantic interoperability, still some major controversies exist, especially with regard to the possibility to achieve uniformity in metadata descriptions and content markup. There is a large group that would work towards a universal ontology that should be applicable in all the SSH domains (and beyond) while others, would rather use pragmatic mappings between parts of different descriptive schema and vocabularies where possible and needed, referring to the huge effort involved in the maintenance of such universal ontologies. In SSHOC the pragmatic approach has been chosen, recommending schema and vocabularies on the basis

[17] http://www.e-rihs.eu
[18] http://www.elra.info
[19] https://www.ldc.upenn.edu
[20] https://www.european-language-grid.eu
[21] https://switchboard.clarin.eu
[22] http://vlo.clarin.eu
[23] https://tei-c.org/release/doc/tei-p5-doc/en/html/

of actual usage, while accepting that others need to work towards new descriptive systems and providing mappings between them. Creating and maintaining specific semantic interoperability solutions (mappings) requires domain expertise and is better done in the context of community organisation collaborations such as SSHOC. However offering a platform that allows easy management and sharing of such mapping solutions can be provided at the EOSC level since such a platform can be domain-agnostic.

5. Concluding remarks and discussion points

The future of LR and LT in the SSH part of EOSC will be partly determined by a number of policies and actions that will be shaped by the many-fold ongoing and envisaged SSHOC activities that are focused on services that can be applied to LR.

- For various reasons it is more effective to have new communities participating in and contributing to EOSC through existing cluster networks such as SSHOC instead of directly via EOSC.
- SSHOC has already successfully demonstrated the potential for sharing and scaling-up infrastructure components.
- SSHOC can play a role in aligning centres from the broader LR community with EOSC, including their industry-oriented activities.
- The desired efficiency in provision of generic services for the SSH community can often be achieved through the collaboration between research community organisations that directly share solutions rather than through the use of services of large service provider organisations.
- In large-scale collaborative projects a pragmatic approach to semantic interoperability solutions is more effective than a single ontology-oriented approaches.
- Especially in the SSH case, where a majority of researchers is less IT-savvy than in the "hard sciences", the communities are better positioned to describe and explain services and solutions.
- SSHOC collaboration will encourage sharing of infrastructure services and components across SSH domains and communities.

In order to achieve the prominent role of the communities both in infrastructure development and maintenance, as is argued and proposed in this position paper, it is paramount that long-term funding schemes and policies are in place to support shared resources and services, and that a proper community-oriented governance layer be set up.

6. Acknowledgements

The work reported here has received funding (through CLARIN ERIC) from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823782 for project SSHOC. We would like to thank Franciska de Jong for providing valuable feedback and engaging the authors in exciting discussions.

7. Bibliographical References

Cieri, C. (2020). Stretching disciplinary boundaries in language resource development and use: a linguistic data consortium position paper.

Hey, A. J. G. and Trefethen, A. E. (2003). The Data Deluge: An e-Science Perspective. In F Berman, et al., editors, *Grid Computing - Making the Global Infrastructure a Reality*, pages 809–824. Wiley and Sons. Chapter: 36.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.