SSHOC
social sciences & humanities open cloud

Collaborative Use Cases between the SSH Open Marketplace, the Language Resource Switchboard and the Virtual Collection Registry

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, Grant Agreement #823782.

Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Collaborative Use Cases between the SSH Open Marketplace, the Language Resource Switchboard and the Virtual Collection Registry

Dissemination Level	PU
Date	05/11/2020
Work Package(s)	WP3 - Lifting Technologies into the SSH Cloud WP7- Creating the SSH Open Marketplace
Task	T3.6 T7.1
Type	Report
Number of Pages	p.1 – p.14

Abstract: The overarching goal of the activity in T3.6 is to achieve - wherever useful and technically possible - an integration between CLARIN and DARIAH components, being developed and extended under the SSHOC umbrella. This includes especially the MP, the VCR and LRS. The DARIAH research infrastructure plays an important role in this regard, although this document focuses on the CLARIN LRS and CLARIN VCR as well as the MP.

Author List

Organisation	Name	Contact Information
CLARIN ERIC	Daan Broeder	daan.broeder@di.huc.knaw.nl
UGOE	Stefan Buddenbohm Marthe Eisner	sbudden@gwdg.de eisner@sub.uni-goettingen.de
ÖAW	Matej Durco Klaus Illmayer	matej.durco@oeaw.ac.at Klaus.Illmayer@oeaw.ac.at

Executive Summary

The CLARIN Language Resource Switchboard (LRS or Switchboard) serves as an established and valuable asset to provide the user with tools and services for their research data. Apart from looking at an already existing set of research data, moreover, many users would like to search for tools and services from the angle of the research method, a certain technology, interoperability or even just a research question, which calls for the SSH Open Marketplace (MP). The MP not only strives to provide individual items (tools, research data, tutorials, software) to the searching and browsing user, but first and foremost context. Contextualized items in the MP allow for a search serendipity which contributes substantially to the service experience. The LRS promises to be a suitable means to convey such serendipity. For this purpose, the following document outlines possible user stories in favor of its integration into the MP.

Beyond the LRS-MP relation the document also considers scenarios for the relation of both components to the CLARIN Virtual Collection Registry (VCR). The VCR allows the user to create individual and persistent collections including records from a broad range of sources. Such sources may be repositories exposing research data (the common VCR use case= 'bibliography of research datasets'), but possibly also from other sources such as the MP.

The overarching goal of the activity in T3.6 is to achieve - wherever useful and technically possible - an integration between CLARIN and DARIAH components, being developed and extended under the SSHOC umbrella. This includes especially the MP, the VCR and LRS. The DARIAH research infrastructure plays an important role in this regard, although this document focuses on the CLARIN LRS and CLARIN VCR as well as the MP.

Content

Aim of this document 6
A - Creation of a Virtual Collection with Records from the MP 7
B - Inclusion of the VCR as Feature in the MP 8
C - Invocation of the LRS from the MP based on MIME Type 9
D - Invocation/Call up of MP from LRS with a Search String 11
E - LRS as Ingested Source in the MP 13
F - MP desktop tool info for the LRS 13
Conclusion 15

Aim of this document

The Social Sciences and Humanities Open Cloud (SSHOC) is the SSH contribution to the European Open Science Cloud (EOSC). An important task in this regard is to investigate possible relations and integration possibilities between CLARIN and DARIAH services to enhance the mutual user experience for the communities.

In concrete terms: What are possible user stories for the integration of the CLARIN LRS and CLARIN VCR into the MP? What could be beneficial from the users' perspective? What are limitations from a technical integration perspective?

This document describes possible scenarios and forms a link between MP and LRS and is also related to a similar user stories document of WP7. We thank the colleagues from CLARIN-D and DARIAH-DE for their critical review and comments to this documents, particularly Thomas Eckart and Nanette Rißler-Pipka.

The LRS serves as an established and valuable asset to assist researchers and data producers with individual tools and services for their research data. Apart from looking at an already existing set of research data, moreover, many users would like to search for tools and services from the angle of the research method, a certain technology, interoperability or even just a research question, which calls for the MP. The MP not only provides individual items (tools, (possibly) research data, tutorials, software) to the searching and browsing user, but first and foremost context. Contextualized items allow for a search serendipity which could contribute substantially to the service experience. The LRS promises to be a suitable means to convey such serendipity.

Beyond the LRS-MP relation the document also considers scenarios for the relation of both components to the VCR. The VCR allows the user to create individual and persistent collections including records from a broad range of sources. Such sources may be repositories exposing research data (the common VCR use case = 'bibliography of research datasets'), but may also include records from other sources, such as the MP.

The document lists user stories for integration scenarios and tries to elaborate the challenges, benefits and feasibility for each of them. The text may serve as a basis for the discussion among WP3 and WP7 in SSHOC and is currently not an implementation roadmap, but merely a collection of concepts.

The overarching goal of the activity in T3.6 is to achieve - wherever useful and technically possible - an integration between CLARIN and DARIAH components, being developed and extended under the SSHOC umbrella. This includes especially the MP, the VCR and LRS. The DARIAH research infrastructure plays an important role in this regard, although this document focuses on the CLARIN LRS and CLARIN VCR as well as the MP.

In this regard, the following aims are pursued by T3.6:

- extend the visibility and use of LRS, VCR as well as MP
- utilise LRS or VCR to provide context on items within the MP
- bond of language resource-affiliated researchers to the MP
- introduce new services to the LRS and overcome the language resources bias
- introduce new data types/MIME types to the LRS
- promote the integration between the SSHOC work packages WP3 and WP7

A - Creation of a Virtual Collection with Records from the MP

This use case describes the use of the VCR to include records from the MP. The virtual collection doesn't have to be entirely composed of MP records, they may form only a subset.

Presentation of content in the MP: The MP provides the user with information on useful resources in the SSH research domains. Usually these resources will be individual items (e.g. tools, tutorials, software) or collections of resources grouped together (e.g. in form of a workflow or research process). Particularly the second scenario seems to be attractive as it offers search serendipity to the user and sets individual items into an explanatory context. However, with this complexity also comes the requirement for adapting and storing these collections.

Scenario: If a researcher finds a useful set of items in the marketplace he or she might be interested in persisting the set as a collection, to adapt to specific needs and to save it for later work. This calls for an integration of the VCR.

What is a collection? The collection can either be composed of pre-fabricated workflows from the MP (similar to the SSK scenarios), which are enhanced by the user OR can be a completely new collection, e.g. a search result list, which gets refined by the user and needs to be stored. An idea for a virtual collection revolves around teaching. Such a virtual collection could serve as a basis for teaching on a specific subject and collect, for instance, text, tools, tutorials, and exemplary datasets.

An important aspect of the VCR affecting also the other user stories is related to its collection approach: One of the services' strengths, but also a weakness of the VCR is that VCs can be overarching different repositories and registries. This implies that the VC metadata needs to remain relatively agnostic with regard to the purpose and idiosyncrasies of a particular repository. For example, if a user wants to describe workflows and to connect this description with data collections in a VC, the metadata should be part of a VC constituent rather than being part of the VC metadata.

Requirements and challenges: Stored collections belong to the user and may also be able to invoke the LRS. In order to realize this, the first option requires an individual user space in the

MP, where these collections can be stored. Due to the current state of development it is not clear, if such individual user spaces will be created. Alternatively, these collections could be stored within the VCR itself or get a PID. The second option requires a linkage between the MP and the LRS. At the current stage it is unclear how the MP resource will be citable. A PID seems to be difficult as the MP's resources will get updated frequently. Nevertheless, it is likely that the researcher, in order to share with colleagues and to cite in own documents, expects the possibility to refer to (at least certain) MP entities. It is not always required that the entities are 'frozen' depending on a versioning policy, the PID could refer to the latest version. This is a technical requirement to be solved on the MP's side.

Benefits: A linkage between the three components (LRS, VCR, MP) seems reasonable because the MP will be an attractive source for creating collections motivated by the quantity and diversity of its content. On the other hand this may motivate users to enhance the marketplace's contents and extend the usage of the LRS and VCR.

Feasibility: Regarding the other user stories, the creation of virtual collections containing MP items seems to be one of the more obvious and useful ones. However, the feasibility for its implementation is dependent on the possibility to reference MP items in a persistent way. This implementation effort has to be yielded on the MP's side.

B - Inclusion of the VCR as Feature in the MP

This use case outlines the integration of the VCR as a feature directly in the MP. Although the use case A covers the user's requirement to create virtual collections containing records from the MP it is possible to think of the VCR as genuine MP feature.

Scenario: This scenario addresses the demand of users to create collections within the MP, including MP records, but also records from other sources.

Requirements and challenges: The implementation effort for this use case relates at least to two main issues: the depth of integration of the VCR in the MP. This should be done in a way that the user experience is as seamless as possible and the user experiences the function as genuine to the MP. The second issue relates to non-MP records. The unique character of the VCR is to gather records from various sources in an actionable way. Without this the virtual collection would be just a conventional list of records. The actionability of an individual record depends on its metadata. As long as the necessary metadata for a record is held by the MP or is available in an easy way through external sources, the user experience will be sufficient.

Benefits: Although it may be possible to create a virtual collection in the VCR directly, which may include MP records it may be a threshold for the user to do so. It is a threshold for the user to leave the MP and switch to a separate web service for the creation of a virtual collection. From a usability point of view it is more convenient to create the collection directly in the MP. Apart from the usability such a function might offer valuable metadata for the MP. Thought from the angle of contextualisation the user created virtual collections might indicate relations between

records that are so far not considered. A user-created collection might also hint to valuable sources that have not been ingested into the MP so far.

Feasibility: The implementation effort of this use case relates to the necessary level of awareness between the two services, MP and VCR. At least three levels of awareness can be identified (coming with different costs):

- If the creation function stays without the MP and is being done in the VCR, reading of the MP API may be sufficient. In this scenario the VCR only retrieves information from the MP through its API and apart from this the whole process and user interaction stays within the VCR. This is covered by use case A.
- More beneficial from a user's point of view would be to create and present the virtual collection within the MP. This requires considerable implementation effort as it should be possible to include sources outside the MP.
- Even more beneficial would be the visibility of these virtual collections within the MP. Then they can be shared with other users, adapted and extended. Beyond the above-mentioned implementation effort the question of the non-MP records arises. In which way can non-MP records be presented if they are part of a virtual collection? The MP won't hold any describing metadata for such records. Displaying such records as wildcards may be dissatisfying for the user but possibly the compromise to implement this use case.

C - Invocation of the LRS from the MP based on MIME Type

This use case describes a possible relation of the MP to the LRS.

If resources in the MP come with a MIME-type - as it should be the case for certain types of research data - it could be conceptually possible to invoke the LRS from the individual resource site of the MP. This invocation could be implemented in a similar way as it had been done with the TextGrid Repository here:

https://hdl.handle.net/11858/00-1734-0000-0005-1421-2

Scenario: The user can quite conveniently click on the 'LRS check button' and gets the instant results from the LRS through a pop-up. The user doesn't leave the MP in this scenario. The displayed results are based on the file's MIME type and will be a selection from the LRS tool inventory:

https://switchboard.clarin.eu/tools

Requirements and challenges: A main challenge will likely be the composition of content in the MP. At the current stage of development the MP aims at tools, tutorials, software, and collections, but not primarily. The described use case relies on the existence of resources coming with certain MIME types. This is a technical requirement resulting from the LRS side as

the identification of presented research data. The subsequent recommendation of selected tools relies on the MIME type. For instance research data with a text related MIME type will lead to the recommendation of text processing tools. To get useful recommendations, this mechanism relies on a well selected tool inventory in the LRS and on the transmission of MIME types, which have to be as specific as possible.

With regard to the MIME type: the MP doesn't need to know the mediatype because LRS has its own mediatype/language detector. Mediatype and language are not mandatory, and in fact currently ignored on LRS side.

Another challenge may be related to the way the MP collects and stores its content. As a reminder: The MP doesn't serve as holder of its content - the resources. The MP just presents describing metadata ON the resources. The resources themselves remain at third party sources, usually the creators or holders. For the LRS invocation this poses a technical problem: To be able to identify the dataset by its MIME type, the LRS needs access to the dataset OR needs the MIME type in the metadata. Both options seem feasible as the MP metadata includes both the MIME type and can include a link to the original dataset.

Benefits: The benefit of this use case is the additional uptake of the LRS by users browsing the MP, who so far are not aware of the LRS. For the MP it may offer an additional function, although not feasible for all types of MP resources. This scenario is beneficial, because it raises awareness for the CLARIN research infrastructure in general. However, currently this may only be applicable for MP users with a research background in language related research topics or linguistics.

Feasibility: An implementation seems feasible with regard to the transfer of metadata between the two services. As soon as it is technically possible to transfer metadata other scenarios may become possible as well. Most likely, the transfer of MIME type would be a good starting point.

The implementation in the MP could be created in two ways: Either staying in the MP environment after the LRS invocation or leaving the MP and turning the user to the LRS. As CLARIN, DARIAH and SSHOC are closely collaborating in lifting services and resources within the cloud, the technical costs for implementation could tip the scale in this use case. Apart from the technical aspect - the transfer of metadata from the MP to the LRS - this scenario requires the existence of suitable MIME types or other types of metadata - describing a relation between two resources - on the MP side.

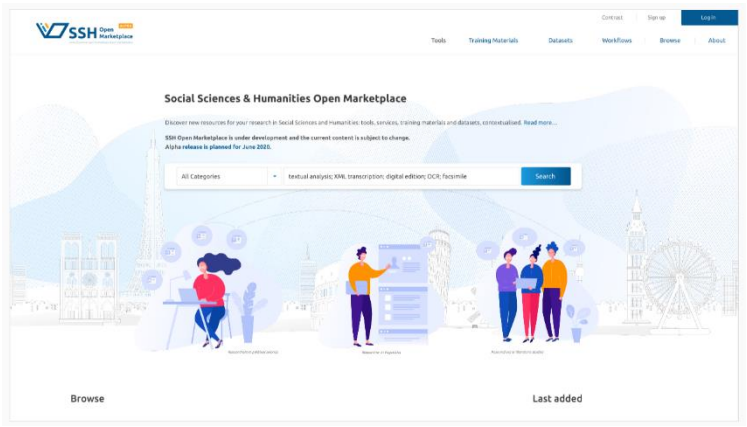

Figure 1: Using the SSH Open Marketplace search slot (Mock up by Justyna Wytrążek)

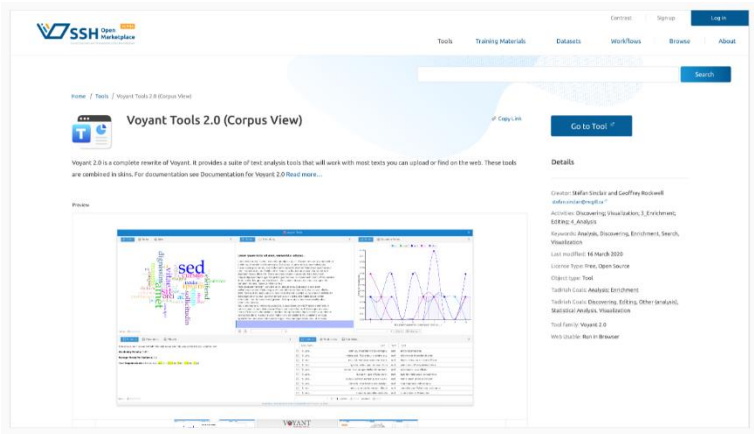

Figure 2: Detailed item view (Mock up by Justyna Wytrążek): on the right side of the item view the VCR invocation or LRS invocation could be implemented.

D - Invocation/Call up of MP from LRS with a Search String

This use case describes an invocation of the MP from the LRS and can be seen as reciprocal to the above mentioned use case B. Technically it is not an invocation, but merely sending a search string (e.g. 'show me similar tools like XY') from the LRS to the MP.

Scenario: In this scenario the user visits the LRS, for instance browses through the tool inventory. The user may have a certain type of research data in mind, but the focus lies on the tools presented by the LRS. Beforehand, the user may also have uploaded a dataset to the LRS and now browses through the list of recommended tools. For example, the user uploaded a TEI file and wants recommendations beyond the available tools of the LRS for this resource type.

In both cases the user finds a call up link at each individual tool within the LRS which could be named 'Show me similar tools at the MP'. Currently it is possible to go through all of the tools at the LRS and compare them manually, but with this invocation of the MP other relevant resources could be presented to the user in an easy way. In this regard it is important to know

that the LRS doesn't hold any user friendly metadata to the tools. The LRS concept relies on the upload of a dataset, identification of MIME type, and/or recommendation of tools. The documentation and description of the tools is not available within the LRS, but usually accessible on the tools' holders websites.

Requirements and challenges: The use case comes with requirements regarding the curation of the MP's resources. Apart from the MIME type mechanism described in use case B, manually curated metadata also seems to be feasible to lead the user to useful recommendations. This could be done on the MP's side with a respective information in the 'relations' sections of the resource.

Both options require at least the transfer of metadata (=the search string e.g. 'show me similar tools') from the LRS to the MP. This requires that both services use a common vocabulary to describe resources. If a presentation of suitable tools within the environment of the LRS is desired, the transfer of metadata also has to be performed back to the LRS. However, this is not necessary for the use case to be functional.

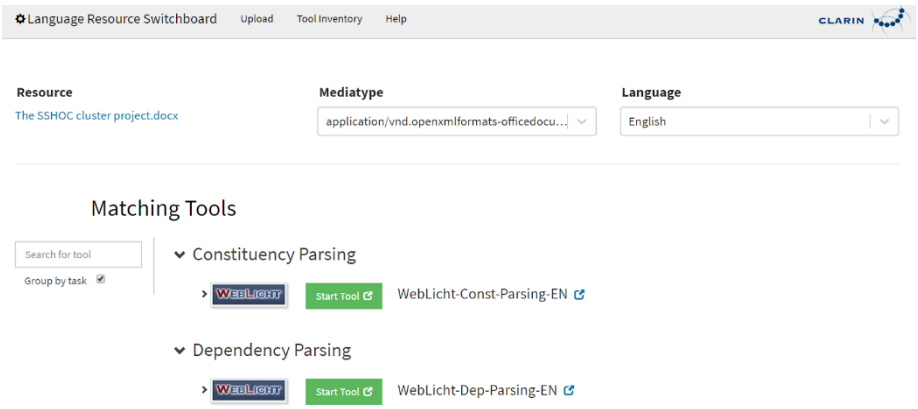

Figure 3: Using the LRS upload function: The function to invoke the MP from the LRS could be implemented as a button similar to "Start Tool" as above. The invocation button should be presented to the user after uploading a research data set and also in the tool inventory.

Benefits: This scenario benefits in raising awareness for the MP. For instance, a researcher of the LRS with a background in linguistics may be introduced to the MP's broad range of resources and may pay off with look at more generic resources, which could be useful for any research discipline. The MP could be seen as a generic extension of the LRS. Where the LRS is currently focussing on language- or linguistic related resources and omitting generic one, this gap could be filled by the MP.

The MP may provide background information with regard to specific services or some complex recipes to further process certain data-types. This information could be made accessible via the Switchboard ie. as a 'help' function.

This use case may also serve as an opportunity to involve the SSHOC interoperability hub of WP3.

Feasibility: The effort for implementation is a technical one - the transfer of metadata between the two services - and the availability of useful metadata on the side of the MP, which is necessary to respond to the incoming request from the LRS.

E - LRS as Ingested Source in the MP

This use case describes the inclusion of the LRS tool inventory on the individual tool level in the MP. The LRS tool inventory is seen as an additional source for ingestion by the MP.

Scenario: In this scenario the user browses the MP for relevant resources with regard to his/her research questions. The user informs himself/herself on tools, tutorials and workflows. Here and there tools from the LRS will appear in the result lists, wherever a feasible relation to the search string of the user is present.

Requirements and challenges: The implementation effort connected to this scenario is overseeable. The LRS tool inventory consists of 73 tools for language related research data (November 2020). This collection has to be ingested into the MP. It remains unclear from where the MP may retrieve the describing metadata for the tools. As said before, the LRS concept is threefold:

1. Users uploads a dataset,
2. LRS identifies the MIME type,
3. LRS suggests tools corresponding with the MIME type.

Any documentation or further metadata on the tools are not delivered by the LRS directly, but directs the user to the websites of the tools' owners.

Benefits: After the ingestion of this collection - which is planned anyway - the LRS tool inventory is visible on a per tool level in the MP. It will also be possible to keep the collection character of the tools. The possibility to browse through the whole LRS tool inventory collection at the MP would most likely improve the user experience, while currently the LRS comes without any further documentation or tutorial regarding the tools (which both are usually available at the tools homepage).

Feasibility: As described above the technical effort to include the LRS tool inventory in the MP is overseeable. The challenge lies in the fact that the tools need descriptions - which is currently not delivered by the LRS directly.

F - MP desktop tool info for the LRS

This use case revolves around desktop based tools, which is a common occurrence in the humanities, but which fall out of the scope of the LRS as it is aimed at web services: Currently the LRS only lists web services and directs the user to them. Desktop based tools are not yet included.

Scenario: This use case is inspired by earlier discussions on the possibility that the MP provides information (sheets) on concepts and entities that are encountered by a user working with the LRS. Until now the best example of such an entity was the data type of a data resource, in which case the user could invoke a link to an information page in the MP. For instance: Explaining the TEI format. In discussions with the German CLARIAH project, it became clear that it is intended that the LRS would also provide information on available desktop installed tools. Even if the user cannot invoke a desktop tool via the LRS, it is useful that the LRS provides information on which desktop tool would be applicable and where to find installation information.

Requirements and challenges: To include desktop installed tools in the MP is possible and a common appearance. For the LRS the case is different as its use concept relies on a web service: The user uploads a research data set and receives a list of recommended tools, which are instantly available to him or her by clicking through. Although this is a very convenient approach for the user it is technically not necessary to leave out a desktop installed tool. The only difference with desktop installed tools in the current LRS concept would be, that the user is not able to click through the service and start right away. Viewed from this end, the inclusion of desktop installed tools in the LRS may be a valuable contribution for the user.

Benefits: At least two goals are met by this use case: Firstly, the LRS widens its current approach to include desktop installed tools as well. Secondly, the user has access to a broader range of tools and might receive more recommendations. This scenario seems very reasonable with regard to the DARIAH offered range of tools - some of them falling into the desktop installed category, but clearly with potential use for users of the LRS.

Feasibility: Currently the LRS doesn't offer any detailed information on the listed tools in the inventory. As long as the listed tools are web services coming with proper documentation on the tool site itself, this is not a problem (although maybe a problem with regard to usability). With desktop tools the question of standardised accompanying metadata arises. Most of these desktop installed tools come with a homepage anyway, but this scenario may offer a good opportunity to rethink the LRS approach to go without standardised metadata.

First it was considered that such desktop tools info would be served from a LRS environment, however it would seem beneficial to both LRS and MP if such information can be provided by the MP and curated in MP. It would also align SSHOC with CLARIAH-DE plans.

Conclusion

The five use cases outlined above describe possible integration scenarios involving the CLARIN Virtual Collection Registry and the CLARIN Language Resources Switchboard on the one hand, and the SSH Open Marketplace on the other hand.

With look at the implementation effort at least the use cases the following ranking results:

1. LRS as ingested source in the MP: already implemented
2. Invocation of VCR from the MP: could be based on MIME type; only applicable for a subset of the MP records
3. Invocation of the MP from LRS: could be facilitated with a search string, which leads to results in the MP; problem of blurred search results which may be not specific enough
4. Creation of a virtual collection within in the VCR with records included in the MP: reliance on the persistent identification of records within the marketplace
5. Inclusion of the VCR as feature in the MP: depending on depth and functionality of integration

One of the scenarios above is already being implemented (inclusion of the tool inventory of the LRS in the marketplace as a subset of records), for the other scenarios the discussion is ongoing. The priority is - without question - on the development of the MP to a productive service and it remains unclear which resources are available for the implementation of additional scenarios described in this document.

Another strang of discussion - but not documented in here - revolves around possible relations of the Virtual Collection Registry, the Language Resource Switchboard on the one hand, and DARIAH resources on the other hand. Such DARIAH resources are the DARIAH Data Federation Architecture, including the Collection Registry and the Data Modeling Environment.

SSHOC
social sciences & humanities open cloud

www.sshopencloud.eu

@SSHOpenCloud

in/company/sshoc

info@sshopencloud.eu

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement **No. 823782**