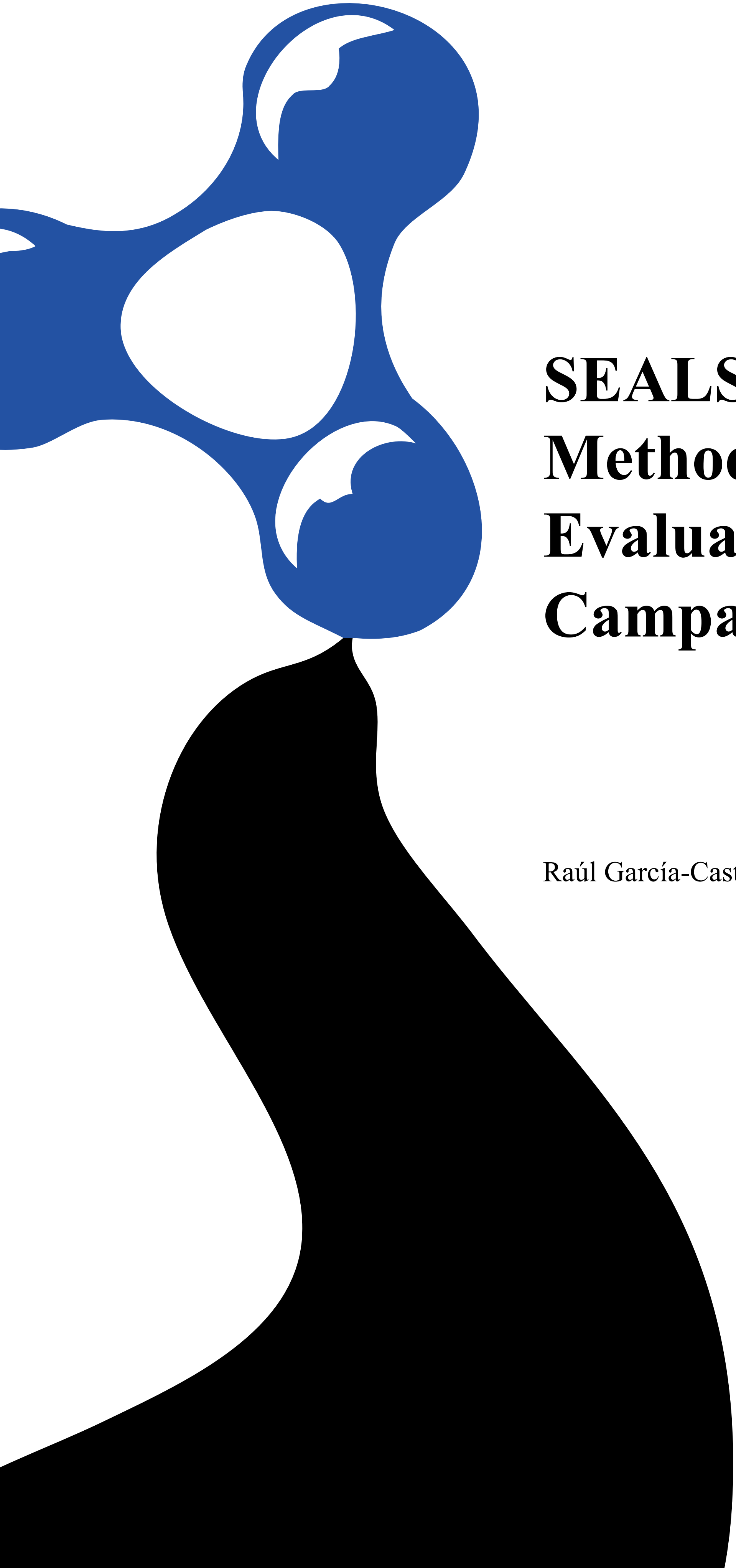

SEVENTH FRAMEWORK
PROGRAMME

SEALS Methodology for Evaluation Campaigns

Raúl García-Castro and Stuart N. Wrigley

September 2011

Document Information

IST Project Number	FP7 – 238975	**Acronym**	SEALS
Full Title	Semantic Evaluation at Large Scale		
Project URL	`http://www.seals-project.eu/`		

Authors	Raúl García-Castro (Universidad Politécnica de Madrid) and Stuart Wrigley (University of Sheffield)			
Contact Author	**Name**	Raúl García-Castro	**E-mail**	rgarcia@fi.upm.es
	Inst.	UPM	**Phone**	+34 91 336 36 70

Abstract	This document describes the second version of the SEALS methodology for evaluation campaigns that is being used in the organization of the different SEALS Evaluation Campaigns.
Keywords	evaluation campaign, methodology, guidelines

Table of Contents

List of Figures

1. Introduction

In the SEALS project, we have issued two evaluation campaigns for each of the different types of semantic technologies covered by the project. In these evaluation campaigns, different tools were evaluated and compared according to a common set of evaluations and test data.

This document describes the second version of the methodology to be followed to implement these evaluation campaigns. The first version of the methodology [1] was defined after analysing previous successful evaluation campaigns and has been improved with the feedback and lessons learnt obtained during the first SEALS Evaluation Campaigns.

This methodology is intended to be general enough to make it applicable to evaluation campaigns that cover any type of technology (either semantic or not) and that could be defined in the future. Because of this, we have described the methodology independently of the concrete details of the SEALS evaluation campaigns to increase its usability. Nevertheless, we also highlight the benefits of following a “SEALS” approach for evaluation campaigns (e.g., automatic execution of evaluations, evaluation resource storage and availability).

The methodology should not be considered a strict step-by-step process; it should be adapted to each particular case if needed and should be treated as a set of guidelines to support evaluation campaigns instead of as a normative reference.

This document describes first, in chapter 2, a generic process for carrying out evaluation campaigns, presenting the actors that participate in it as well as the detailed sequence of tasks to be performed. Then, chapter 3 includes the agreements that rule the SEALS evaluation campaigns and that can be reused or adapted to other campaigns.

2. Evaluation campaign process

This chapter presents a process to guide the organization and execution of evaluation campaigns over software technologies. An evaluation campaign is an activity where several software technologies are compared along one or several evaluation scenarios, i.e., evaluations where technologies are evaluated according to a certain evaluation procedure and using common test data.

A first version of this process appeared in [1] and has been the one followed in the first round of evaluation campaigns that have been performed in the SEALS project. This document updates that previous version including feedback from people organising and participating in those evaluation campaigns.

First, the chapter presents the different actors that participate in the evaluation campaign process; then, a detailed description of such process is included, including the tasks to be performed and different alternatives and recommendations for carrying out them.

2.1 Actors

The tasks of the evaluation process are carried out by different actors according to the kind of roles that must be performed in each task. This section presents the different kind of actors involved in the evaluation campaign process.

- **Evaluation Campaign Organizers** (Organizers from now on). The Organizers are in charge of the general organization and monitoring of the evaluation campaign and of the organization of the evaluation scenarios that are performed in the evaluation campaign. Depending on the size or the complexity of the evaluation campaign, the Organizers group could be divided into smaller groups (e.g., groups that include the people in charge of individual evaluation scenarios).

- **Evaluation Campaign Participants** (Participants from now on). The Participants are tool providers or people with the permission of tool providers that participate with a tool in the evaluation campaign.

SEALS value-added features

In SEALS there is a committee, the Evaluation Campaign Advisory Committee (formerly named the Evaluation Campaign Organizing Committee), in charge of supervising all the evaluation campaigns performed in the project to ensure that they align to the SEALS community goals. This committee is composed of the SEALS Executive Project Management Board, the SEALS work package leaders and other prominent external people. Each evaluation campaign is led by different groups of Evaluation Campaign Organizers (formerly named the Evaluation Campaign Executing Committee) who coordinate with the Evaluation Campaign Advisory Committee.

2.2 Process

This section describes the process to follow for carrying out evaluation campaigns. Since this process must be general to accommodate different types of evaluation campaigns, this methodology only suggests a set of general tasks to follow, not imposing any restrictions or specific details in purpose.

Some tasks have alternative paths that can be followed. In these cases, the methodology does not impose any alternative but presents all the possibilities so the people carrying out the task can decide which path to follow.

The description of the tasks is completed with a set of recommendations extracted from the analysis of other evaluation campaigns, as presented in [1]. These recommendations are not compulsory to follow, but support specific aspects of the evaluation campaign.

Figure 2.1 shows the main phases of the evaluation campaign process.

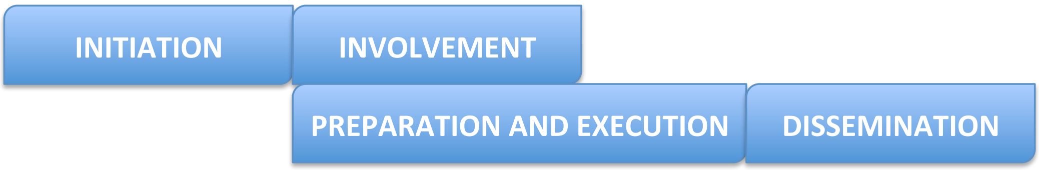

Figure 2.1: The evaluation campaign process.

The evaluation campaign process is composed of four phases, namely, Initiation, Involvement, Preparation and Execution, and Dissemination. The main goals of these phases are the following:

- **Initiation phase**. It comprises the set of tasks where the different people involved in the organization of the evaluation campaign are identified and where the different evaluation scenarios are defined.
- **Involvement phase**. It comprises the set of tasks in which the evaluation campaign is announced and participants show their interest in participating by registering for the evaluation campaign.
- **Preparation and Execution phase**. It comprises the set of tasks that must be performed to insert the participating tools into the evaluation infrastructure and to execute each of the evaluation scenarios and analyse their results.
- **Dissemination phase**. It comprises the set of tasks that must be performed to disseminate the evaluation campaign results and to make all the evaluation campaign results and resources available.

These four phases of the evaluation campaign process are described in the following sections, where a definition of the tasks that constitute them, the actors that perform these tasks, its inputs, and its outputs are provided.

2.2.1 Initiation phase

The *Initiation* phase comprises the set of tasks where the different people involved in the organization of the evaluation campaign and the evaluation scenarios are identified and where the different evaluation scenarios are defined. These tasks and their interdependencies, shown in figure 2.2, are the following:

1. Identify organizers.
2. Define evaluation scenarios.

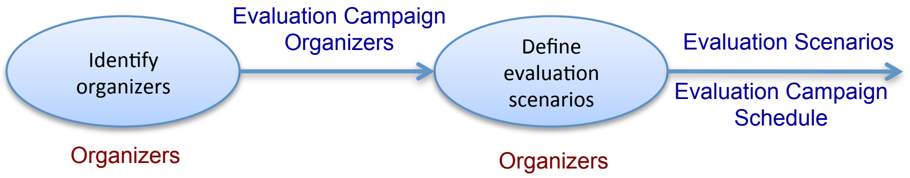
Identify organizers
Organizers
Evaluation Campaign Organizers
Define evaluation scenarios
Organizers
Evaluation Scenarios
Evaluation Campaign Schedule

Figure 2.2: *Initiation* phase of the evaluation campaign process.

Identify organizers

Actors	
E. C. Organizers	
Inputs	**Outputs**
	E. C. Organizers

The goal of this task is to define the group of people who will be in charge of the general organization and monitoring of the evaluation campaign as well as of organizing the evaluation scenarios to be performed in the evaluation campaign and of taking them to a successful end.

Recommendations for the evaluation campaign	
Requirement	**Recommendation**
+ objectiveness	The evaluation campaign does not favor one participant over another.
+ consensus	The decisions and outcomes in the evaluation campaign are consensual.
+ transparency	The actual state of the evaluation campaign can be known by anyone.
+ participation	The organization overhead of the evaluation campaign is minimal.

Recommendations for the evaluation campaign organization	
Requirement	**Recommendation**
+ credibility	The evaluation campaign is organized by several organizations.
+ openness	Organizing the evaluation campaign is open to anyone interested. Organizing an evaluation scenario is open to anyone interested.
+ relevance	Community experts are involved in the organization of the evaluation campaign.

SEALS value-added features

Different people in the SEALS community have expertise in organizing evaluation campaigns for different types of semantic technologies. Don't hesitate to ask for advice or collaboration when organizing an evaluation campaign!

Define evaluation scenarios

Actors	
E. C. Organizers	
Inputs	**Outputs**
E. C. Organizers	Evaluation Scenarios Evaluation Campaign Schedule

In this task, the Organizers must define the different evaluation scenarios that will take place in the evaluation campaign and the schedule to follow in the rest of the evaluation campaign.

For each of these evaluation scenarios, the Organizers must provide the complete description of the evaluation scenario; we encourage to follow the conventions proposed by the ISO/IEC 14598 standard on software evaluation [2].

Evaluation scenario descriptions should at least include:

- Evaluation goals, quality characteristics covered and applicability (i.e., which requirements tools must satisfy to be evaluated).
- Test data (i.e., evaluation inputs), evaluation outputs, metrics, and interpretations.
- Result interpretation (i.e., how to interpret and visualize evaluation results).
- Evaluation procedure and the resources needed in this procedure (e.g., software, hardware, human).

Three different types of test data can be used in an evaluation scenario:

- **Development test data** are test data that can be used when developing a tool.
- **Training test data** are test data that can be used for training a tool before performing an evaluation.

- **Evaluation test data** are the test data that are used for performing an evaluation.

Recommendations for evaluation scenarios	
Requirement	**Recommendation**
+ credibility	Evaluation scenarios are organized by several organizations.
+ relevance	Evaluation scenarios are relevant to real-world tasks.
+ consensus	Evaluation scenarios are defined by consensus.
+ objectiveness	Evaluation scenarios can be executed with different test data. Evaluation scenarios do not favor one participant over another.
+ participation	Evaluation scenarios are automated as much as possible.

Recommendations for evaluation descriptions	
Requirement	**Recommendation**
+ transparency	Evaluation descriptions are publicly available.
+ participation	Evaluation descriptions are documented.

Recommendations for test data	
Requirement	**Recommendation**
+ objectiveness	Development, training and evaluation test data are disjoint. Test data have the same characteristics. Test data do not favor one tool over another.
+ consensus	Test data are defined by consensus.
+ participation	Test data are defined using a common format. Test data are annotated in a simple way. Test data are documented.
+ transparency	Test data are reusable. Test data are publicly available.
+ relevance	Test data are novel. The quality of test data is higher than that of existing test data.

SEALS value-added features

The SEALS Platform already contains different resources for the evaluation of semantic technologies (i.e., evaluation workflows, services and test data). These resources are publicly available so they can be reused to avoid starting your evaluation from scratch.

2.2.2 Involvement phase

The *Involvement* phase comprises the set of tasks in which the evaluation campaign is announced and participants show their interest in participating by registering for the evaluation campaign. These tasks and their interdependencies, shown in figure 2.3, are the following:

1. Announce evaluation campaign.
2. Provide registration mechanisms.
3. Participant registration.

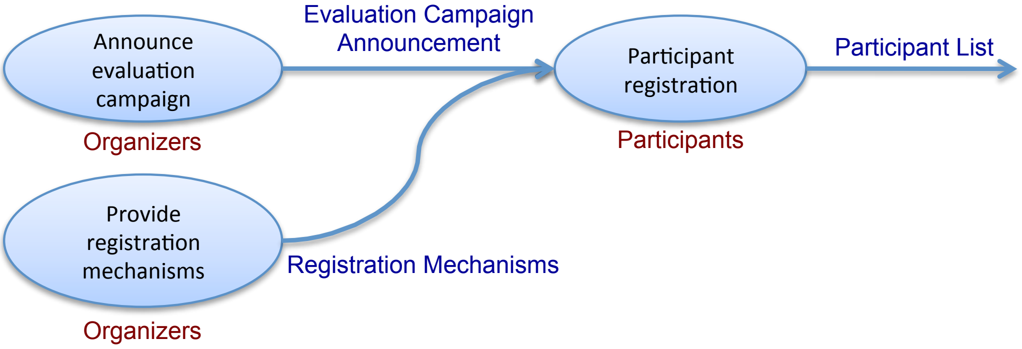

Figure 2.3: *Involvement* phase of the evaluation campaign process.

Announce evaluation campaign

Actors	
E. C. Organizers	
Inputs	**Outputs**
Evaluation Scenarios Evaluation Campaign Schedule	Evaluation Campaign Announcement

Once the different evaluation scenarios are defined, the Organizers must announce the evaluation campaign using any mechanism available (e.g., mailing lists, blogs, etc.) with the goal of reaching the developers of the tools that are targeted by the evaluation scenarios.

Recommendations for evaluation campaign announcement	
Requirement	**Recommendation**
+ participation	The evaluation campaign is announced internationally. Tool developers are directly contacted.

Recommendations for evaluation campaign participation	
Requirement	**Recommendation**
+ openness	Participation in the evaluation campaign is open to any organization.
+ participation	The effort needed for participating in the evaluation campaign is minimal. People can participate in the evaluation campaign regardless of their location. Participation in the evaluation campaign does not require attending to any location. The requirements for tool participation are minimal. Participants have time to participate in several evaluation scenarios.
+ relevance	Community experts participate in the evaluation campaign. Tools participating in the evaluation campaign include the most relevant tools.

SEALS value-added features

The SEALS community dissemination services (e.g., SEALS Portal, mailing lists, blog, etc.) can support spreading your evaluation campaign announcements to a whole community of users and providers interested in semantic technology evaluation.

Provide registration mechanisms

Actors	
E. C. Organizers	
Inputs	**Outputs**
Evaluation Scenarios	Registration Mechanisms

In this task, the Organizers must provide the mechanisms needed to allow potential participants to register and to provide detailed information about themselves and their tools.

Recommendations for registration mechanisms	
Requirement	**Recommendation**
+ participation	Participants can register in the evaluation campaign regardless of their location.

SEALS value-added features

If you don't want to implement your own registration mechanisms, the SEALS Portal can take care of user registration for your evaluation campaign.

Participant registration

Actors	
E. C. Participants	
Inputs	**Outputs**
Evaluation Campaign Announcement Registration Mechanisms	Participant list

In this task, any individual or organization willing to participate in any of the evaluation scenarios of the evaluation campaign must register to indicate their willingness to do so.

> ***SEALS value-added features***
>
> If you are using the SEALS Platform for executing your evaluation campaign, participant registration through the SEALS Portal can be coupled with different evaluation services (e.g., uploading the participant tool into the platform).

2.2.3 Preparation and execution phase

The *Preparation and execution* phase comprises the set of tasks that must be performed to insert the participating tools into the evaluation infrastructure, to execute each of the evaluation scenarios, and to analyse the evaluation results. The tasks that compose this phase can be performed either independently for each evaluation scenario or covering all the evaluation scenarios in each task. These tasks and its interdependencies, shown in figure 2.4, are the following:

1. Provide evaluation materials.

2. Insert tools.

3. Perform evaluation.

4. Analyse results.

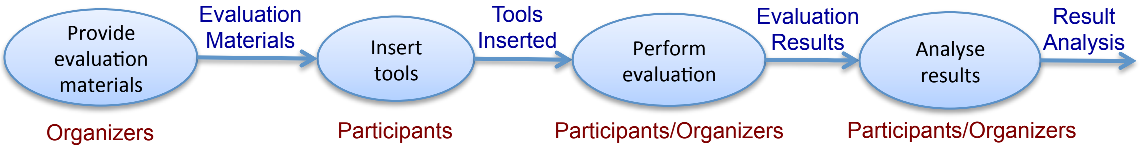

Figure 2.4: *Preparation and execution* phase of the evaluation campaign process.

Provide evaluation materials

Actors	
E. C. Organizers	
Inputs	**Outputs**
Evaluation Scenarios	Evaluation Materials

In this task the Organizers must provide to the registered participants all the evaluation materials needed in the evaluation, including:

- Instructions on how to participate.
- Evaluation description.
- Evaluation test data.
- Evaluation infrastructure.
- Any software needed for the evaluation.

Alternatives:

⊗ *Evaluation test data availability*

a. Evaluation test data are available to the participants in this task.

b. Evaluation test data are sequestered and are not available to the participants.

⊗ *Reference tool*

a. Participants are provided with a reference tool and with the results for this tool. This tool could be made of modules so participants don't have to implement a whole tool to participate.

Recommendations for evaluation material provision	
Requirement	**Recommendation**
+ objectiveness	All the participants are provided with the same evaluation materials.
+ participation	Participants are provided with all the necessary documentation.

Recommendations for evaluation software	
Requirement	**Recommendation**
+ openness	The software used in the evaluation is open source.
+ transparency	The software used in the evaluation is publicly available.
+ participation	The software used in the evaluation is robust and efficient.

SEALS value-added features

Using the SEALS Platform participants can access all the evaluation materials stored in the platform repositories either manually through the SEALS Portal or automatically using the SEALS Platform services.

Insert tools

Actors	
E. C. Participants	
Inputs	**Outputs**
Evaluation Materials	Tools Inserted

Once the Participants have all the evaluation materials, they must insert their tools into the evaluation infrastructure and ensure that these tools are ready for the evaluation execution.

Recommendations for tool insertion

Requirement	**Recommendation**
+ participation	Participants can use test data to prepare their tools. Participants can check that their tool provides the outputs required. Participants can test the insertion of their tools. The insertion of tools is simple.

SEALS value-added features

Similarly to participant registration, tool insertion can be managed through the SEALS Portal. Furthermore, the SEALS Platform can check that tools are correctly inserted so they are ready for execution.

Perform evaluation

Actors	
E. C. Participants/E. C. Organizers	
Inputs	**Outputs**
Tools Inserted	Evaluation Results

In this task, the evaluation is executed over all the participating tools and the evaluation results of all the tools are collected.

Alternatives:

⊗ *Evaluation execution*

a. The execution of the evaluation is performed by the Organizers.
b. The execution of the evaluation is performed by the Participants.
c. The execution of the evaluation is performed by the Organizers and the Participants.

⊗ *Time for executing evaluations*

a. Participants have a limited period of time to return their evaluation results.

⊗ *Use of auxiliary resources*

a. Participants cannot use any auxiliary resource in the evaluation (development and training test data, external resources, etc.).

Recommendations for evaluation execution	
Requirement	**Recommendation**
+ participation	Evaluation execution requires few resources. Evaluation execution can be made regardless of the location or time.
+ objectiveness	All the tools are evaluated following the same evaluation description. All the tools are evaluated using the same test data.
+ credibility	The results of the evaluation execution are validated.

SEALS value-added features

If all the resources needed in an evaluation (i.e., evaluation workflow, test data and tools) are stored inside the SEALS Platform, the platform can automatically execute your evaluation and store all the produced results.

Analyse results

Actors	
E. C. Participants/E. C. Organizers	
Inputs	**Outputs**
Evaluation Results	Result Analysis

Once the evaluation results of all the tools are collected, they are analysed both individually for each tool and globally including all the tools.

This results analysis must be reviewed in order to get agreed conclusions. Therefore, if the results are analysed by the Organizers then this analysis must be reviewed by the Participants and vice versa, that is, if the results are analysed by the Participants they must be reviewed by the Organizers.

Alternatives:

⊗ *Analysis of evaluation results*

a. The analysis of the evaluation results is performed by the Organizers.
b. The analysis of the evaluation results is performed by the Participants.
c. The analysis of the evaluation results is performed by the Organizers and the Participants.

⊗ *Anonymised results*

a. The results of the evaluation campaign are anonymised so the name of the tool that produces them is not known.

Recommendations for evaluation results

Requirement	Recommendation
+ objectiveness	Evaluation results are analysed identically for all the tools. Participants can comment on the evaluation results of their tools before making them public. There is enough time for analysing the evaluation results.
+ transparency	Evaluation results are publicly available at the end of the evaluation campaign. Evaluation results can be replicated by anyone.
+ sustainability	Evaluation results are compiled, documented, and expressed in a common format.

SEALS value-added features

Evaluation results stored in the SEALS Platform can be accessed either through the SEALS Portal by means of visualization services or automatically through the platform services so they can be used in your own calculations. Besides, you can easily combine the results of different evaluations to suit your goals.

2.2.4 Dissemination phase

The *Dissemination* phase comprises the set of tasks that must be performed to disseminate the evaluation campaign results and to make all the evaluation campaign result and resources available. The tasks that compose this phase can be performed either independently for each evaluation scenario or covering all the evaluation scenarios in each task. These tasks and its interdependencies, shown in figure 2.5, are the following:

1. Disseminate results.

2. Publish outcomes.

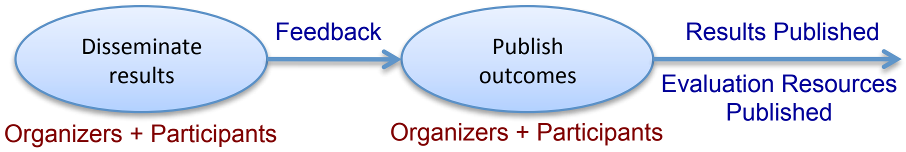
Disseminate results
Feedback
Publish outcomes
Results Published
Evaluation Resources Published
Organizers + Participants
Organizers + Participants

Figure 2.5: *Dissemination* phase of the evaluation campaign process.

Disseminate results

Actors	
E. C. Organizers E. C. Participants	
Inputs	**Outputs**
Result Analysis	Feedback

In this task, the Organizers and the Participants must disseminate the results of the evaluation campaign. The preferred way of dissemination is one that maximizes discussion to obtain feedback about the evaluation campaign (e.g., a face-to-face meeting with participants or a workshop).

Recommendations for result presentation	
Requirement	**Recommendation**
+ consensus	Evaluation results are presented and discussed in a meeting.
+ transparency	Participants write the description of their tools and their results.
+ objectiveness	The presentation of results is identical for all the tools. The presentation of results includes the reasons for obtaining these results.

Recommendations for the dissemination meeting	
Requirement	**Recommendation**
+ openness	The dissemination meeting is public.
+ relevance	The dissemination meeting is collocated with a relevant event.
+ participation	Participants present the details and results of their tool.
+ credibility	The difficulties faced during the evaluation campaign are presented.
+ consensus	Feedback about the evaluation campaign is obtained.
+ sustainability	The dissemination meeting includes the discussion of the next steps to follow. A survey is performed to organizers, participants and attendants.

SEALS value-added features

The SEALS community dissemination services (e.g., SEALS Portal, mailing lists, blog, etc.) can support disseminating your evaluation campaign results and attracting people to meetings where they can be discussed.

Publish outcomes

Actors	
E. C. Organizers E. C. Participants	
Inputs	**Outputs**
Feedback	Results and Resources Published

In this task, the Organizers and the Participants are encouraged to document the results of the evaluation campaign and of each of the tools, including not only the results obtained but also improvement recommendations based on the feedback obtained and the lessons learnt during the evaluation campaign.

Finally, the Organizers must make publicly available all the evaluation resources used in the evaluation campaign, including any resources that were not made available to participants (e.g., sequestered evaluation test data).

Recommendations for result publication	
Requirement	**Recommendation**
+ participation	Results are jointly published by all participants (e.g., as workshop proceedings, journal special issues, etc.).

Recommendations for resource publication	
Requirement	**Recommendation**
+ transparency	All the evaluation resources are made publicly available.
+ sustainability	Test data are maintained by an association.

SEALS value-added features

Making evaluation resources public once the evaluation campaign is over is straightforward since all those resources are already available from the SEALS Platform. Furthermore, the SEALS Portal is the best place to upload or link to any report resulting from your evaluation campaign.

3. Evaluation campaign agreements

This chapter presents the general terms for participation in the SEALS evaluation campaigns and the policies for using the resources and results produced in these evaluation campaigns. They are based on the data policies of the Ontology Alignment Evaluation Initiative[1].

These terms for participation and policies are only a suggestion. They are being used in the SEALS evaluation campaigns but can be adapted as seen fit for other campaigns.

3.1 Terms of participation

By submitting a tool and/or its results to a SEALS evaluation campaign the participants grant their permission for the publication of the tool results on the SEALS web site and for their use for scientific purposes (e.g., as a basis for experiments).

In return, it is expected that the provenance of these results is correctly and duly acknowledged.

3.2 Use rights

In order to avoid any inadequate use of the data provided by the SEALS evaluation campaigns, we make clear the following rules of use of these data.

It is the responsibility of the user of the data to ensure that the authors of the results are properly acknowledged, unless these data are used in an anonymous aggregated way. In the case of participant results, an appropriate acknowledgement is the mention of this participant and a citation of a paper from the participants (e.g., the paper detailing their participation). The specific conditions under which the results have been produced should not be misrepresented (an explicit link to their source in the SEALS web site should be made).

These rules apply to any publication mentioning these results. In addition, specific rules below also apply to particular types of use of the data.

Rule applying to the non-public use of the data

Anyone can freely use the evaluations, test data and evaluation results for evaluating and improving their tools and methods.

Rules applying to evaluation campaign participants

The participants of some evaluation campaign can publish the results as long as they cite the source of the evaluations and in which evaluation campaign they were obtained.

Participants can compare their results with other published results on the SEALS web site as long as they also:

[1] http://oaei.ontologymatching.org/doc/oaei-deontology.html

- compare with the results of all the participants of the same evaluation scenario; and
- compare with all the test data of this evaluation scenario.

Of course, participants can mention their participation in the evaluation campaign.

Rules applying to people who did not participate in an evaluation campaign

People who did not participate in an evaluation campaign can publish their results as long as they cite the sources of the evaluations and in which evaluation campaign they were obtained and they need to make clear that they did not participate in the official evaluation campaign.

They can compare their results with other published results on the SEALS web site as long as they:

- cite the source of the evaluations and in which evaluation campaign they were obtained;
- compare with the results of all the participants of the same evaluation scenario; and
- compare with all the test data of this evaluation scenario.

They cannot pretend having executed the evaluation in the same conditions as the participants. Furthermore, given that evaluation results change over time, it is not ethical to compare one tool against old results; one should always make comparisons with the state of the art. In the case that this comparison is not possible, results can be compared with older results but it must be made clear the age of the result and the version of the tool that produced them.

Rules applying to other cases

Anyone can mention the evaluations and evaluation campaigns for discussing them.

Any other use of these evaluations and their results is not authorized (you can ask for permission however to the contact point) and failing to comply to the requirements above is considered as unethical.

References

[1] R. García-Castro and F. Martín-Recuerda. D3.1 SEALS Methodology for Evaluation Campaigns v1. Technical report, SEALS Project, October 2009.

[2] ISO/IEC. *ISO/IEC 14598-6: Software product evaluation - Part 6: Documentation of evaluation modules.* 2001.