E-RIHS
EUROPEAN RESEARCH INFRASTRUCTURE FOR HERITAGE SCIENCE

E-RIHS PP

CALL: H2020-INFRADEV-2016-2

TYPE OF ACTION: CSA

GA n.739503

D.10.2

Heritage Hub online

Lead Author: Elisabetta Andreassi

With contributions from: Laura Benassi

Deliverable nature	Report (R)
Dissemination level	Public
Contractual delivery date	31 July 2018
Actual delivery date	
Version	1.0
Total page of number	
Keywords	Heritage Science, Hub

Abstract

WP10 is currently carrying out dissemination and communication activities, based on a participatory and inclusive approach and tailored dissemination strategies to communities of users and stakeholders (through workshops, roundtables, conferences, symposia or open days), in order to assure a presentation of E-RIHS to citizens, researchers, investors, private and public bodies, institutions, academic universities or research centres.
Digital (website; sets of digital content and hardware to be used at events) and physical (posters, templates for presentations, flyers, leaflets, press releases) channels of communication are being tested in a dialogue with national E-RIHS representatives and National Hubs as well as with users.
The domain www.e-rihs.eu is the main managing, communication and dissemination platform. The website provides links to E-RIHS social media profiles and national hubs websites.
A further step being carried out in the framework of T10.3 is to include a Digital Hub for Research and Cooperation in Heritage Science, acting as an access point for information, services, good practices, training and virtual meetings for stakeholders to discuss topics of interest. It will link to related initiatives, such as http://heritageportal.eu/.
In order to reach this goal, a call for tender was launched by CNR, task leader of T10.3, for the design and development of a semantic cloud service, SCHEME (Semantic Content retrieval engine for the Heritage hub Empowerment), composed of two functionalities: a semantic web tool and a collaboration platform.
By using semantic technologies, SCHEME intends to provide a selection of content automatically taken from the web and filtered in order to present only the information related to the main themes of E-RIHS. It will be useful to gather external data and optimize content and it will offer a social media analysis platform, to search for relevant posts and identify most used concepts, hashtags, web pages and sites cited in social discussions, geographical location of users, influencers, etc.
By creating a collaboration platform, called the Heritage Hub, E-RIHS will have an easy-to-use but powerful collaboration platform able to act as an access point for information, services, good practices, training and virtual meetings for researchers, users, stakeholders to discuss topics of interest.

Document information

Project number	739503	**Acronym**	E-RIHS PP
Full title	European Research Infrastructure for Heritage Science – Preparatory Phase		
Project url	www.e-rihs.eu		
Document url			
EU Project Officer	Maria Theofilatou		

Deliverable	**Number**	D.10.2	**Title**	Heritage Hub online
Work Package	**Number**	10	**Title**	Advocacy, Communication and Dissemination

Date of delivery	**Contractual**	M18	**Actual**	
Status	Version 1.0		☑ Final	
Nature	❑ Prototype X report ❑ demonstrator ❑ other			
Dissemination level	X Public ❑ restricted			

Authors (Partner)	CNR			
Responsible Author	**Name**	Elisabetta Andreassi	**Email**	elisabetta.andreassi@cnr.it
	Partner	CNR	**Phone**	0832321816
Responsible Author	**Name**	Laura Benassi	**Email**	laura.benassi@ino.cnr.it
	Partner	CNR	**Phone**	0552308221

Abstract (for dissemination)	
Keywords	Heritage Science, Hub

Version Log			
Issue Date	**Rev. no.**	**Author**	**Change**
00/00/0000	0.1		

Table of contents

List of figures

Abbreviations

Narrative (technical) description

Introduction

Due to its global dimension, E-RIHS aims at offering itself as a digital Hub for Research and Cooperation in the field of Heritage Science. The idea was to design a semantic web tool acting as an access point for information, services, good practices, training and virtual meetings for discussing topics of interest.
The Heritage Hub is designed as a cloud service for sharing specific and multiform interests and information in an effective way. By using semantic technologies, it also provides a selection of content automatically taken from the web and filtered in order to present only the information related to the main themes of the project.

Call for tender: conditions and description

In April 2018, CNR as the task leader launched a call for tender for the design and development of the semantic web tool, from now on called SCHEME, whose acronym means Semantic Content retrieval engine for the Heritage hub Empowerment.

For the design of the tool, three macro-phases have been identified:

1. Beta version of the SCRE and THH platform (phase 1): Within 90 days from the contract agreement
2. Final version of the SCRE and THH platform (phase 2): Within 120 days from the definition of the final version
3. Maintanance and helpdesk (phase 3): Until January 31th, 2020

Detailed description of the tool

The tool is designed for a double purpose: on the one hand it allows the automatic retrieval, curation and reuse (e.g. on the Infrastructure site) of web content fetched from multiple sources and identified as “interesting” through semantic and NLP algorithms; on the other it is used as a tool to empower the collective work of social communities through the automatic retrieval and sharing of valuable content, gathered from the web and related to topics of interest, and the provision of collaboration and networking tools.
SCHEME is composed of two functionalities:

1. a Semantic Content Retrieval Engine (SCRE),
2. a collaboration platform, named The Heritage Hub (THH).

By using semantic technologies, SCHEME will provide a selection of content automatically taken from the web and filtered in order to present only the information related to the main themes of E-RIHS. It will be useful to gather external data and optimize content and it will offer a social media analysis platform, to search for relevant posts and identify most used concepts, hashtags, web pages and sites cited in social discussions, geographical location of users, influencers, etc.

By creating a collaboration platform, called the Heritage Hub, E-RIHS will have an easy-to-use but powerful collaboration platform able to act as an access point for information, services, good practices, training and virtual meetings for researchers, users, stakeholders to discuss topics of interest.

Motivation and main goals

SCHEME must be implemented around two main functionalities:
These two services are strongly related, since the automatic retrieval, selection and curation of content of interest will be the main asset through which collaboration is intended to function for the communities involved in the Infrastructure. Herein we describe in details these two services and how they must be integrated to provide the maximum collaboration experience for E-RIHS users.
All SCHEME components (SCRE and THH) must be designed as cloud services, completely accessible, both for their front-end and back-end interfaces, with a modern and standard web browser.
All softwares must be developed preferably by using a whole stack of open source components and libraries: the use of Open Source software will be considered an extra value and it will represent one of the voice ("Extra conditions") to be evaluated in the technical proposal. The use of specific commercial cloud services (e.g. for semantic analysis) is admitted but must be clearly justified.
SCHEME will be installed on servers, using a high availability strategy, provided by E-RIHS.

Specific functionalities

The provision of a powerful and engaging collaboration platform (The Heritage Hub/THH) is the core of the present project. It is fundamental to point out that collaboration is centered in SCHEME around topics of interest: in order to better engage users and to encourage them to visit frequently THH, it is strategic to provide valuable contents, linked to the broad nature of E-RIHS activities. By searching the Internet, one can find a continuous flux of articles of interest published on the websites of research centres, scientific magazines, online newspapers etc; this, together with the ever increasing amount of User Generated Content in Social Networks and personal blogs, represents a goldmine of valuable content that can be republished in THH to feed discussions amongst researchers.
While this strategy is certainly interesting and might represent a practical solution to encourage recurring visits to THH, it is also significantly expensive and slow to be implemented, if all research and content curation is done manually. Paramount to the success of this vision is the ability to automatize and make configurable the whole process of scanning the web in search of content, by selecting it through a semantic engine based on its main topics and by cleaning its HTML code, in order to ease its re-use/publication, on THH and on other websites. Manual intervention should be therefore limited to a minimum: once configuration has been applied, the engine must work seamlessly by searching, selecting and returning valuable articles and posts from the Internet.
The Semantic Content Retrieval Engine (SCRE) is the enabling component in SCHEME that must allow performing automatic and semi-automatic curation of web content, fetched from multiple sources and identified as items of interest through semantic and NLP algorithms.
SCRE must not be implemented as a content management system but should act as a "smart mediator" for the destination platforms and websites (THH to start with), where content will be published. Fetched content must be provided to the destination platforms via APIs, e.g. standard syndication protocols like RSS or ATOM or through REST web services.

SCRE must be highly configurable through a web dashboard, in order to easily allow both the addition of new interesting sources to be monitored and the definition of semantic rules that identify whether or not a piece of content can be of interest for the destination platforms.
When a piece of content is fetched and considered "valuable", it must be also pre-processed in a way that allows its seamless integration on the destination sites. All decorative elements from the source page (e.g. header, footer, navigation bars and menus, breadcrumbs, advertisement, related articles,...) must be automatically removed, in order to return an HTML code that is as clean as possible: this way, publishing on the destination platforms does not disrupt their presentation and layout.
SCRE must be also designed as a multi-tenant cloud service, in order to allow its integration on different destination platforms. For each one of them, it must be possible to configure individual and independent rules for fetching, filtering and reusing content.
THH on the other hand must be designed as an easy to use but powerful collaboration platform. By mimicking the interactions of people in social networks, it must allow to easily build communities of researchers sharing specific and multiform interests and freely exchanging information in an agile but effective way.
As said, in THH collaboration will revolve around contents of interest that will be mostly acquired automatically through the SCRE service: its semantic features will allow presenting only information related to the main themes of the project. The vision is to stimulate discussions and exchange of ideas around topics of interest. THH therefore must provide services to allow partners to communicate and collaborate, include web forums, document management services, personal messaging, etc. These tools will be also used to promote and collect feedback to ensure the project is tracking the reality of the evolving needs of the communities and the point of view of all stakeholders.

Technical Requirements for SCRE

SCRE must be seen as a smart mediation engine, that allows gathering contents of interest from the web: it should not provide any front-end interface since publishing must be completely based on "output channels", that is an API layer, based on standard syndication protocols (RSS, ATOM) or JSON/REST web services.
From a functional point of view ,the system must support the following features:

Configurable publishing rules

SCRE must provide a configurable logic for the whole publishing workflow. In particular, through the configuration dashboard, operators can:

- define the "sources" of content;
- define the "output channels", each one with a unique ID, that provide selected content for destination sites;
- associate sources to output channels;
- define for each output channel a set of semantic rules (see below) to specify whether content should be considered for publishing or discarded.

In particular SCRE must support the retrieval of content from at least the following sources:

- Web sites through crawling/spidering;
- Syndication feed, via standard protocols (RSS, ATOM);
- Twitter, with all the possible options that its APIs provide (e.g. search of content through keywords, mentions or hashtags, retrieval of posts from single accounts, geographical search, etc);
- Facebook through all the possible options that its APIs provide;

Publishing must be based on API keys in order to ensure that SCRE can only be used by authorized parties.

Semantic rules and classification mechanism

Semantic Rules allow specifying whether or not a piece of content, retrieved from a source, must be maintained and associated to an output channel, or discarded. Their definition can be based at least on the following selection mechanisms:

- Textual keywords, including hashtags, that the fetched text must contain;
- Associated concepts, taken from a widely used and all-compassing controlled dictionary (e.g. an "upper ontology" or a general purpose collection like Wikipedia). Each concept can have associated a weight: in this case the relevancy of the retrieved piece of content is measured by how its text presents concepts that are semantically similar/near to those defined in the rule, by using semantic and NLP algorithms. It does not necessarily work on a direct correspondence: a post can satisfy a rule even if it does not directly contain the concepts defined in it but some other entity that is somewhat (and more or less directly) in association with them. The weight mechanism should allow defining stricter or looser associations between the concepts specified in the rule and those identified in the text;
- Language filtering rules, to allow the selection of posts based on the languages in which they are written. English, French and Italian must be supported, while the availability of filters for German, Spanish, and other languages.

SCRE must also provide classification mechanisms, based on configurable taxonomies. In particular, the service must allow the automatic classification of the retrieved content based on the entries of these ad-hoc taxonomies.
SCRE taxonomies must be also shared with THH.

Multitenancy
SCRE must support independent working groups, where each one of them can define their own publishing rules without interfering with the configurations of the others. It must support the feature through specific user management functionalities that also support role-based access control mechanisms.

Description of the configuration workflow on SCRE

SCRE must be a highly configurable platform: while powerful in features, its back-end interfaces are designed with usability in mind. In particular, it is important to provide a direct feedback to the operator for every operation performed on the platform.
The steps for configuring SCRE can be described as envisioned in the sequence that follows:

- An operator creates an output channel for a specific initiative (e.g. a THH Virtual Folder: see below): the back-end interface provides the details of the endpoint for retrieving the channel content.
- The operator firstly selects the possible web sources for the web content. Secondly, the operator defines the type of the source (e.g. a web site) and all the relevant parameters to configure the content retrieval process (e.g. the URL of the site, the depth of the crawling process, the frequency of the acquisition, etc. A "sample" of the retrieved content can be immediately visualized: if it is ok, the source is saved and associated to the channel. This process is repeated for every interesting source to be analyzed (e.g. a Twitter search). As the operator adds the sources, an immediate feedback of the composition of the channel is visible.
- The operator defines the semantic rules for the output channel. This can be done by:
    - defining language filters, that allow fetching only content in a specified language
    - specifying keywords that the text must include
    - defining a semantic rule with a list of topics that the content must include, directly or indirectly (e.g. with similar concepts) in its text.

Again, once the rules have been specified, the operator should have an immediate feedback of how they apply on the content of the current channel.

- The operator builds a taxonomy as a hierarchy of categories and defines for each one of them a semantic rule based on a list of concepts and an associated weight. If the retrieved content satisfies the rule, it will be automatically classified with the corresponding category.

Technical Requirements for THH

THH must be designed as a user collaboration platform, allowing a community to share specific and multiform interests, to fully and successfully communicate and interact.
THH must allow users to share interests, information and content. The creation of relationships revolves around areas of common interests: all THH functionalities must be designed to ease communication and to allow the exchange of information between users in an agile and friendly way.
The functionalities that the platform must provide are:

- Users must be able to join the platform by directly registering on it or importing their profiles from the most common Social Networks (e.g. Google+ or Facebook). Single sign-on with the aforementioned Social Networks must be also ensured.
- Virtual folders: to ease the grouping of related activities, a customizable and general classification mechanism must be provided. Each content and service of THH can be associated to a Virtual Folder (VF).
- SCRE integration: streams from content channels exposed by SCRE must be imported automatically. They can be also automatically associated to a VF. The main taxonomies used in THH must be synchronised with those defined in the SCRE service.
- General Content Dashboard (GCD), where articles taken from the web through SCRE (but also by editing content directly in THH) will be published. Articles can be classified thematically, allowing THH operators to provide extra categories to those automatically proposed by SCRE. They can be also associated to a VF.
- Personalised Content Dashboard. It allows users to create a personal collection of the articles published in the GCD: the association can be either performed manually (that is, the user explicitly selects the articles of interest) or can be based on simplified semantic rules. The term “simplified” here implies that the process of defining the rule must be intuitive and easy to perform for the final user. Finally, the user must have a personal and private classification mechanism, based on simple mechanisms like tags, to better arrange the content of collection.
- VF Stream: it provides a quick up-to-date view of all content and activities classified for the specific VF.
- Web Forums to allow users to be engaged in online discussions through sequences of messages and related answers: it should be possible to comment or “like” each post. Web Forums can be associated to a specific VF.
- Cloud-based file-system service, where documents (Office files, text and PDF documents, multimedia content like images) can be uploaded and shared with other users. It should be also possible to create and share a Wiki Page, that can be directly created and edited via a web browser.
- Search: all content published on THH, can be searched, by using both a simple and an advanced interface (including text-based search and faceted filtering). It must be also possible to find users, both by their personal information and by concepts, that is by specifying their interests, either defined explicitly in their profiles or "inferred" implicitly according to their activities on the hub platform.
- Messages: to allow direct and private communication between users.
- Advanced collaboration features on content, e.g. the possibility to take and share notes directly on web pages of the THH.

At the beginning of the work, the E-RIHS communication team and the company met to discuss the preliminary scheme of the tool. The E-RIHS team supervised the activity and helped the company define the initial configuration set-up of the SCHEME system and configure rules for SCRE and THH.

References

It will link to related initiatives, such as http://heritageportal.eu/.

It will provide a set of advanced services – including a dedicated social tool for innovation in cultural heritage, developed in the IPERION CH project – that allow partners to communicate and collaborate: discussion forums, document management, collaborative wikis, polls and surveys, content import from RSS feeds, personal messaging.

This will be a powerful tool for the Editorial Committee of E-RIHS to easily produce edited content for publication on the Hub itself. The Hub will be on- line before M18 (D10.2).