Distributed Metadata Schema and Demonstrator for Open Humanities Methods

Claudia Engelhardt, Claudio Leone, Yoann Moranville

Humanities at Scale

D8.1 Distributed Metadata Schema and Demonstrator for Open Humanities Methods

HaS-DARIAH
INFRADEV-3-2015-Individual implementation and operation of ESFRI projects
Grant Agreement no.: 675570

Date: 30-10-2017
Version: 1.0

Project funded under the Horizon 2020 Programme

Grant Agreement no.:	675570
Programme:	Horizon 2020
Project acronym:	HaS-DARIAH
Project full title:	Humanities at Scale: Evolving the DARIAH ERIC
Partners:	DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN – KNAW GEORG-AUGUST-UNIVERSITAET GOETTINGEN STIFTUNG OEFFENTLICHEN RECHTS
Topic:	INFRADEV-3-2015
Project Start Date:	01-09-2015
Project Duration:	28 months
Title of the document:	D8.1 Distributed Metadata Schema and Demonstrator for Open Humanities Methods
Work Package title:	Open Methods Infrastructure
Estimated delivery date:	31-10-2017
Lead Beneficiary:	UGOE-SUB
Author(s):	Claudia Engelhardt [claudia.engelhardt@sub.uni-goettingen.de] Claudio Leone [claudio.leone@sub.uni-goettingen.de] Yoann Moranville [yoann.moranville@dariah.eu]
Quality Assessor(s):	Franziska Helbing [helbing@sub.uni-goettingen.de]
Keywords:	open data, open humanities data platform, research tools, research service, research infrastructure, research registry

Revision History

Version	Date	Author	Beneficiary	Description
0.1	05-04-2017	Claudia Engelhardt	UGOE-SUB	First draft
0.2	21-07-2017	Claudia Engelhardt Yoann Moranville	UGOE-SUB DARIAH	Second draft
0.3	20-10-2017	Claudia Engelhardt Claudio Leone Yoann Moranville	UGOE-SUB UGOE-SUB DARIAH	Third draft
1.0	30-10-2017	Claudia Engelhardt Claudio Leone Yoann Moranville	UGOE-SUB UGOE-SUB DARIAH	Final report

Table of Content

Executive Summary

The goal of HaS task 8.1 according to the description of work was to develop a "distributed metadata schema" that can be used to embed machine-readable descriptions of tools and services in community-driven websites (hence distributed) and a "demonstrator for open humanities methods", i.e. a registry that collects this information and makes it accessible via a central point of entry. Previous community efforts were included in both of these subtasks. The Metadata Application Profile draws on the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH) and the NeDiMAH Methods Ontology (NeMO), among others. For the demonstrator, the Tools E-Registry for E-Social science, Arts and Humanities (TERESAH), originally created by the EU project DASISH[1] that ended in December 2014, was re-used and adapted.
About the nature of this document: This deliverable is the first one of work package 8 "Open Methods Infrastructure". It is a short description of the work done within this work package and its results: the HaS Metadata Application Profile and the TERESAH demonstrator[2].

	Nature of the deliverable	
	R	Document, report
	DEM	Demonstrator, pilot, prototype
	DEC	Websites, patent fillings, videos, etc.
✓	OTHER	
	Dissemination level	
✓	P	Public
	CO	Confidential only for members of the consortium (including the Commission Services)
	EU-RES	Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)
	EU-CON	Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)
	EU-SEC	Classified Information: SECRET UE (Commission Decision 2005/444/EC)

Disclaimer

The Humanities at Scale is project funded by the European Commission under the Horizon 2020 programme. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

[1] Data Service Infrastructure for the Social Sciences and Humanities, http://dasish.eu/
[2] http://teresah.dariah.eu/

1. Introduction

In digital humanities, research methods highly depend on the tools and services used to put them into practice. One cannot make use of a specific method without a tool at hand that supports it and, of course, understanding how it works. Without knowledge about the tools' mode of operation, it is also difficult to validate or replicate research results. In order to establish an effective and sustainable infrastructure for open arts and humanities, it is not enough only to promote openness in terms of access to publications and data, but also with regard to methods, tools and services.
The project Humanities at Scale (HaS)[3] - which is aiming to advance DARIAH by growing the DARIAH community, developing core services and informing research communities - took up this challenge as part of its work. In work package 8 "Open methods infrastructure", it addresses two aspects of this topic area. The first one is the discussion of DH methods, the tools and services used with them and the dissemination of this knowledge. The second area concerns the collection and provision of information about which tools exist and where one can get a hold of them as well as a basic description of their functionalities and operation mode. This report deals with the second topic, worked on in task 8.1 of said work package, and documents the activities of the HaS project in this respect.

The goal of task 8.1 was to develop solutions for a more sustainable description of tools and services, namely a Metadata Application Profile (AP) and a registry demonstrator. In doing so, HaS focuses on a decentral approach in which the descriptions are not entered directly into the registry, but are implemented in the websites of projects using or providing a tool or service in machine-readable form utilising RDFa. The descriptions will then be harvested and displayed by the registry. In addition to that, entries can also be made manually.

The AP and registry will benefit both researchers and providers of tools and services.
They will help researchers

- to find tools or services that support a certain method more easily
- make their research (that uses a certain tool or service) more visible in an easy way (by enhancing its findability for search engines).

Tool and service providers will be able to

- make their tool or service more visible in an easy way.

List of online resources linked to the deliverable:

- Demonstrator TERESAH:
  http://teresah.dariah.eu/
- User manual:
  http://teresah.dariah.eu/help

[3] http://has.dariah.eu/

- GitHub source code and documentation:
  https://github.com/DARIAH-ERIC/TERESAH/

2. State of the art: Registries and metadata for DH tools and services

2.1. Existing tools and service registries

To first get an overview of the current situation, particularly in Europe, the National Coordinators of the DARIAH members were asked for input on tools and service registries and metadata used by these in their respective countries in June 2016. About half of the persons asked responded. The answers indicated that there were no such registries in the DARIAH member countries at that date.[4]
For Italy and the Netherlands, the respondents reported developments towards tools and service registries. In the Netherlands, efforts in this direction were still in the planning stage at the time of the survey. A registry aimed not at tools and services specifically but at projects already available in the beta version was the Dutch Digital Humanities Registry.[5] It is organised around projects which are classified into types - and one of these types is "Tool". So by drilling down the search to that, one can find a number of tools developed or provided by Dutch DH projects.
The Italian registry was in development at the time and has gone live in the meantime.[6] It contains information on tools and services provided by DARIAH-IT. Each entry has a description which usually also contains a URL. The search can be drilled down by providing institutions, scholarly domain and resource type.
In some countries there are lists - mostly in the form of text documents - of existing DARIAH tools and services, for example in Croatia and Greece. No designated metadata schemas were used in these.

In addition, desk research on the topic was undertaken. The result was similar to that of the survey: There were hardly any registries for tools and services specifically targeted at humanities research. However, we found two registries that could be relevant in the context of task 8.2: the Digital Research Tools (DiRT) Directory[7] and the Tools E-Registry for E-Social Science, Arts and Humanities (TERESAH).[8]
The DiRT Directory is a registry of Digital Research Tools for scholarly use. Its origins lie in the Project Bamboo[9] (2008-2012), "a cyberinfrastructure initiative for the arts and

[4] One respondent noted that, for smaller countries with a DH community not that huge, the need for a registry might not have emerged yet.
[5] Website: http://dh-projectregistry.org/
[6] Website: http://it.dariah.eu/sito/strumenti/
[7] Website: https://dirtdirectory.org/
[8] On GitHub: https://github.com/DASISH/TERESAH
[9] See https://dirtdirectory.org/about

humanities"[10]. Therefore, it has a focus on arts and humanities - but it is open to generic tools or tools from other domains as well.

The metadata guidelines provided by DiRT[11] specify twenty fields, five of which are mandatory. Three of the fields relate to TaDiRAH (see 2.2.).

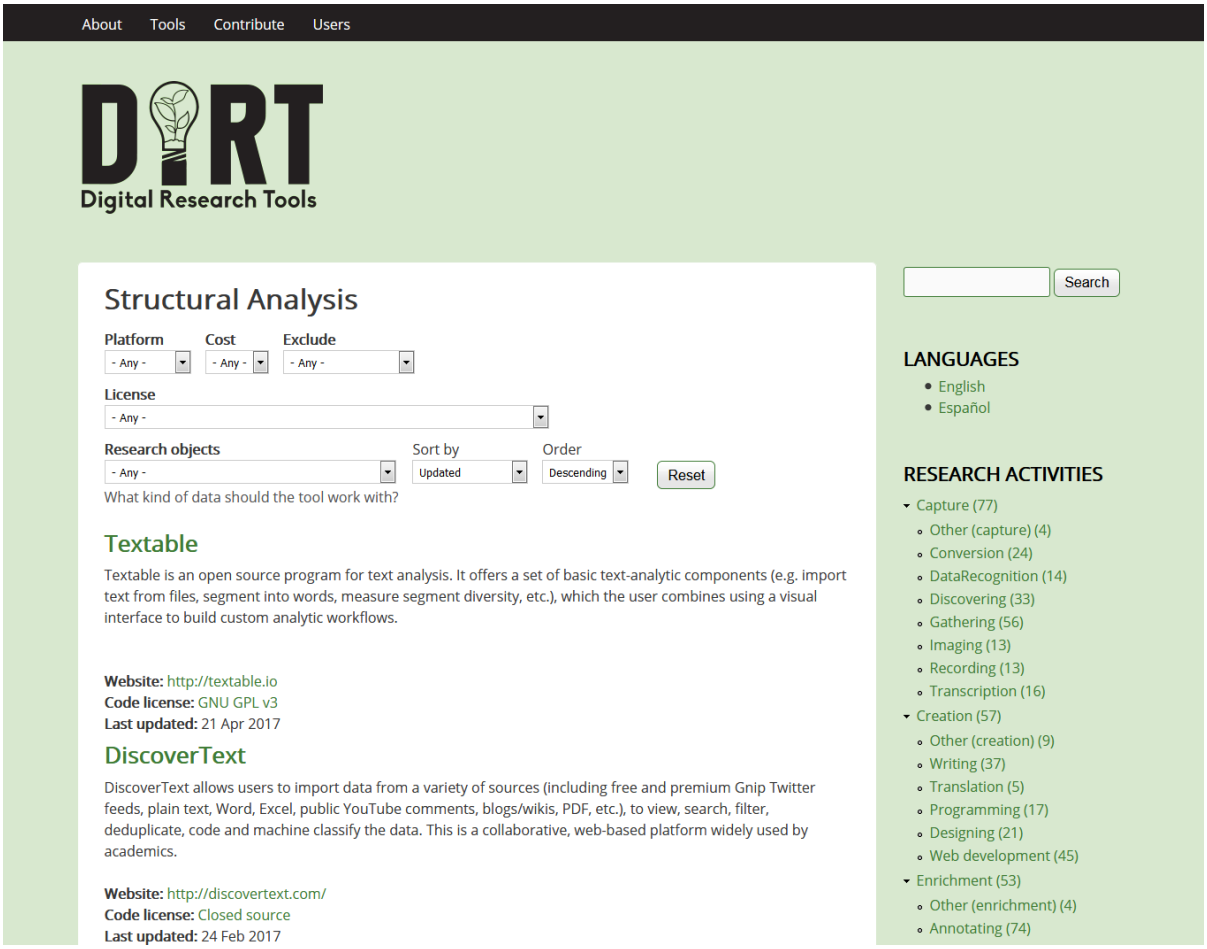

Figure 1: The Digital Research Tools Registry (DiRT), here: search by category view

TERESAH, the Tools E-Registry for E-Social Science, Arts and Humanities, was developed by DASISH (Data Service Infrastructure for the Social Sciences and Humanities)[12], a project funded under the Seventh Framework Programme of the EU that ran from 2012 to 2014. DASISH aimed to bring together the five ESFRI infrastructures from the Social Sciences and Humanities (CESSDA, CLARIN, DARIAH, ESS and SHARE) and develop solutions for common problems and demands in the four areas data quality, data archiving, data access and legal and ethical aspects.[13] In this respect, it is similar to

[10] http://www.projectbamboo.org/
[11] DiRT Metadata Guidelines https://docs.google.com/spreadsheets/d/1Z8-b-LkchyvrmTQHI1GMsJNEWp2IEqLNePhGj-47PXY/edit#gid=1279384626
[12] Website: http://dasish.eu/
[13] See http://dasish.eu/about_dasish/mission/

Humanities at Scale that is working on common solutions with a focus on DARIAH-EU and its seventeen individual members. TERESAH can integrate data from external sources - by the time of the release it did so with data from DiRT. Another way to create new entries is to do it manually. At the time of its release, TERESAH contained 676 tools. The metadata are not as extensive as the ones in DiRT: name (the only mandatory field), developer, keyword, license, platform, standard and tool type.

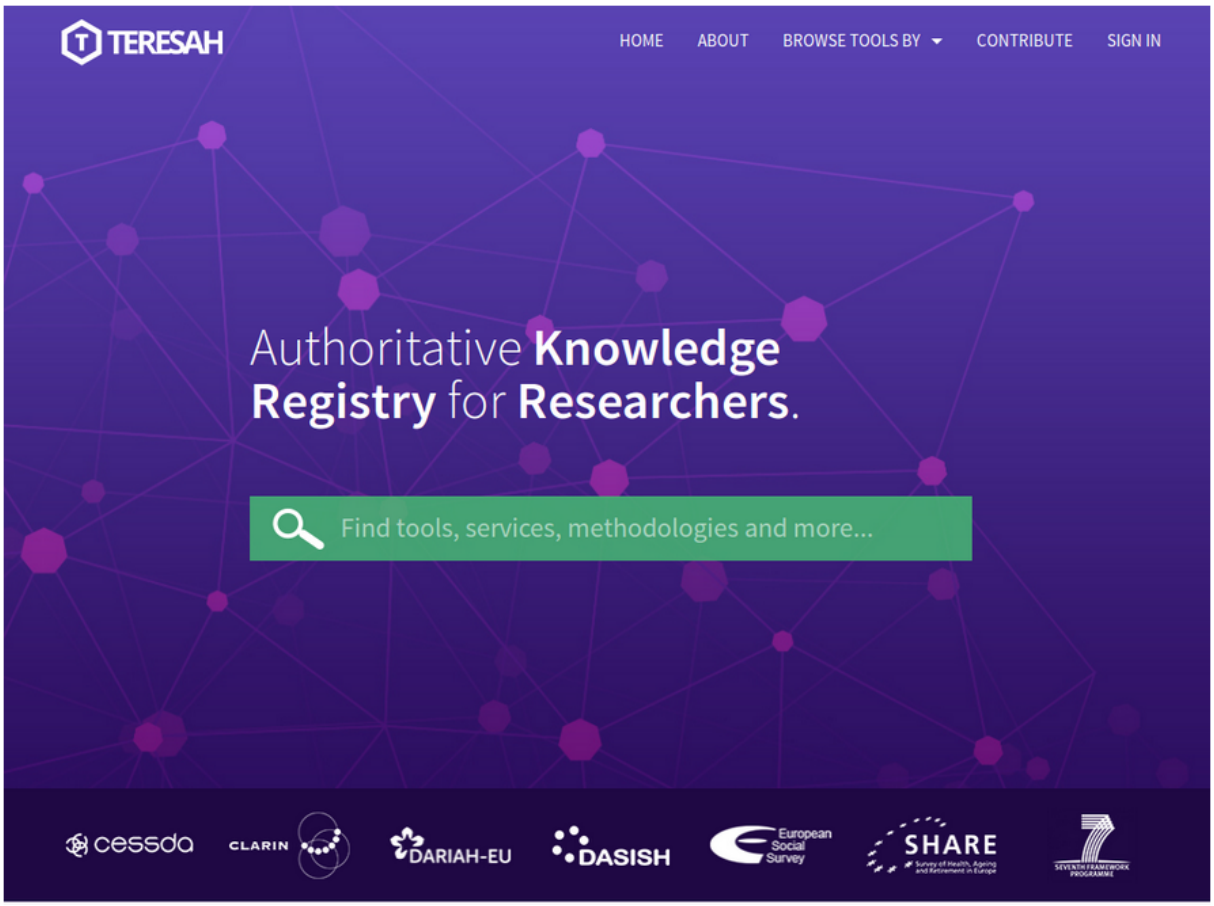

Figure 2: The TERESAH front page when it was released in 2014 (screenshot taken from the TERESAH user manual on GitHub)

2.2. Elements used for the development of the Metadata Application Profile

To make tools and services findable, they - like any other object - have to be described in an appropriate way. In library and information science, the terms used for this purpose are called metadata (also often referred to by their literal meaning as "data about data"). The more accurate and detailed the metadata are, the easier the described objects are to discover in the vastness of the world wide web. Quality, consistency and interoperability are important aspects in this respect. To ensure these, a variety of standards have been

developed for different contexts.[14] According to Boughida[15], metadata standards can be classified into four types:[16]

- **Data structure standards**
  Data structure standards can be described as "'categories' or 'containers' of data that make up a record or other information object". Typically, these are metadata schemas or metadata element sets. Examples are the Dublin Core Element Set, the MARC (Machine-Readable Cataloging) Format or EAD (Encoded Archival Description).
- **Data value standards**
  Data value standards "are used to populate data structure standards or metadata element sets". They are controlled vocabularies, thesauri, controlled lists etc. used to describe the objects in question. They are usually topic- or subject-specific. Examples from the Digital Humanities are TaDiRAH (Taxonomy of Digital Research Activities in the Humanities) and NeMO (NeDIMAH Methods Ontology) which will be explained in more detail later.
- **Data content standards**
  Data content standards are "guidelines for the format and syntax of the data values that are used to populate metadata elements", e.g. codes or cataloguing rules.
- **Data format / technical interchange standards**
  These are often versions of data structure standards in machine-readable form, e.g. RDF (Resource Description Framework), MARC 21 or Simple Dublin Core XML.

One can combine components of several standards into an **AP** that is tailored to a certain community or use case. In addition to elements from different element sets, an AP can also contain locally defined elements. The important thing is, that in its entirety it fulfils the functional requirements of the specific use case.[17]

We identified a number of metadata standards, controlled vocabularies and related standards that provide the necessary elements for the description of tools and services used in arts and humanities research. They are briefly described in the following.

[14] See Gilliland, Anne J. "Setting the Stage." In: Introduction to Metadata, edited by Murtha Baca. 3rd ed. Los Angeles: Getty Publications, 2016. http://www.getty.edu/publications/intrometadata/setting-the-stage/
[15] See Boughida, Karim B.: "CDWA lite for Cataloguing Cultural Objects (CCO): A new XML schema for the cultural heritage community." In: Proceedings of the XVI international conference of the Association for History and Computing, 14-17 September 2005, Royal Netherlands Academy of Arts and Sciences Amsterdam, 2005, pp. 49-56. http://www.dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANShumanitiescomputersandculturalheritageUK.pdf
[16] The following explanations rely on Table 1 in Gilliland, Anne J. 2016. All quotations in the bullet points are also taken from there.
[17] See Baca, Murtha. Baca, Murtha. "Glossary ." In Introduction to Metadata, edited by Murtha Baca. 3rd ed. Los Angeles: Getty Publications, 2016. http://www.getty.edu/publications/intrometadata/glossary/

2.2.1. Metadata standards

Schema.org[18]

Schema.org is a collaborative effort initiated by Internet companies that operate search engines "to create, maintain, and promote schemas for structured data on the Internet"[19]. It can be used in combination with RDFa, Microdata or similar formats to mark up information on a website and make it machine-readable. This greatly enhances the findability of the tagged content by search engines.[20]

The schema.org vocabulary consists of types which are arranged in a hierarchical order. Each type has a related set of properties. To date, the core vocabulary comprises 597 types and 875 properties.[21]

Schema.org can be classified as a data structure standard. Because its purpose - create structured data to enhance the findability of contents on the internet - is in accord with the goal of task 8.1, it was selected as the main data structure standard for the AP. Terms from schema.org in combination with Dublin Core terms make up 20 of 22 properties of the AP developed by HaS.

Dublin Core

Dublin Core (often referred to in its abbreviated form DC) is a metadata standard for the description of digital resources on the web. Its development dates back to the mid-1990s and was particularly driven by libraries and other cultural heritage institutions.[22]

There is a "basic" set of fifteen more or less generic properties: the Dublin Core Metadata Element Set, Version 1.1.[23] These provide the cornerstone for the description of a large number of different kinds of resources. They are embedded in a larger set of vocabularies, the DCMI Metadata Terms[24], which can be used to add further terms to accommodate specific requirements. The terms of the Element Set are: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type.

Dublin Core is a well-established data structure standard, especially in the cultural heritage domain. It was used together with schema.org to provide the property terms for the AP.

2.2.2. Ontologies and vocabularies used

To assign information or content to a metadata field or property, either free text can be used or controlled vocabularies. Where possible, the use of controlled vocabularies

[18] Website: http://schema.org/
[19] http://schema.org/
[20] See http://schema.org/docs/gs.html
[21] See http://schema.org/docs/schemas.html
[22] See https://de.wikipedia.org/wiki/Dublin_Core
[23] Website: http://dublincore.org/documents/dces/
[24] Website: http://dublincore.org/documents/dcmi-terms/

should be preferred, because they allow a distinct designation of the content. In the AP, this could not be realised for all or not even the majority of fields. This is due to the fact that for a considerable number of the properties in the AP there simply do not exist applicable vocabularies. With TaDiRAH there is an instrument available which is specifically tailored to the digital humanities. Its three subsets were applied in association with properties describing how and for what purpose a certain tool can be employed in digital humanities research. Another digital humanities-specific instrument that was used is the NeDiMAH methods ontology (NeMO).

Taxonomy of Digital Research Activities in the Humanities (TaDiRAH)[25]

TaDiRAH is a taxonomy aimed at mapping the field of Digital Humanities by classifying and categorising the respective activities. Its development was a joined effort by the team of the DiRT directory, who were seeking to revise the ad-hoc set of categories which the directory originally used, and a DARIAH-DE working group intending to further develop an overview of digital humanities methods and procedures they had already assembled previously. The team followed a bottom-up approach based on existing taxonomies.[26]

TaDiRAH consists of three sets of elements:

- Research Activities
- Research Objects
- Research Techniques.

The Research Activities list is a closed list consisting of two tiers. The first one comprises eight categories that relate to research goals: seven referring to phases or components of the research process (Capture, Creation, Enrichment, Analysis, Interpretation, Storage, Dissemination) plus one representing "Meta-Activities" like community building or project management. On the more fine-grained tier underneath, each of these categories contains three to seven methods that can be employed to attain the respective research goal. The other two sets are open lists of digital research objects and research techniques respectively. Used in combination, the three sets allow describing different facets of, for example, a tool.[27]

In terms of the types of standards described above, TaDiRAH can be classified as a data value standard. In the AP, its three sets are used as controlled vocabularies the properties *Application Category* (Research Activities), *Is Used For* (Research Techniques) and *Research Object* (Research Object).

[25] Website: http://tadirah.dariah.eu/vocab/index.php

[26] See Borek, Luise; Dombrowski, Quinn; Perkins, Jody; Schöch, Christoph. "TaDiRAH: a Case Study in Pragmatic Classification. In: Digital Humanities Quarterly. Vol. 10, Nr. 1 2016. http://digitalhumanities.org/dhq/vol/10/1/000235/000235.html

[27] See ibid.

NeMO
The NeDiMAH Methods Ontology (NeMO) "is a comprehensive ontological model of scholarly practice in the arts and humanities"[28], developed by the ESF Research Network NeDIMAH (Network for Digital Methods in the Arts and Humanities). It consists of 29 classes[29] that are arranged in a hierarchical order of several tiers. The three superclasses on the highest level are Activity, Actor and Object.[30] Each class has a number of properties. NeMO has a special focus on the interrelation between entities of these three groups.
In the AP, elements of NeMO were used as a vocabulary for the property "Type" (the subclasses of NeMO class "Instrument") and as a property (Is Used For).

As data content standards, ISO 8601 or W3CDTF are suggested for the standardised expression of dates for the properties "DateCreated" and "DateModified".
RDFa is employed as a data format / technical interchange standard to express the metadata in machine-readable form. More on that in chapter 4.

3. Metadata Application Profile

The AP was developed in an iterative process. In August and September 2016, the WP8 team assembled a first working document with a list of 29 potential properties. The document was circulated internally among the SUB and FHP colleagues involved in HaS. The feedback included a ranking of the properties, because originally, in the final version the number of properties should, if possible, be limited to around 15 in order to keep the effort for implementing it manageable. Together with the incorporation of feedback on other aspects, this resulted in a first draft. The limit of 15 could not entirely be kept, but mandatory and non-mandatory properties were introduced to set the threshold not too high for potential users wishing to implement it.
This first draft was presented to and discussed with the HaS consortium at the Project Board Meeting in Ghent on 12-13 October 2016. The overall feedback was very positive. Some minor adaptations were proposed and comments made. For example, we first envisioned VIAF[31] to be used as controlled vocabulary for the properties "Provider" and "Creator", but HaS colleagues pointed out that not every institution or person will be in there, so this was dropped subsequently. Also, "ContactPerson" was proposed as an additional property (but not adopted).[32] After considering and incorporating the feedback, a first version was finalised.

[28] http://nemo.dcu.gr/
[29] See http://nemo.dcu.gr/index.php?p=navigate, click on "Classes view".
[30] See http://nemo.dcu.gr/index.php?p=navigate, click on "Graph view".
[31] https://viaf.org/
[32] This idea was considered, but eventually dropped due to the fact that indicating a concrete person always bears the risk that the information will become invalid sooner or later, because people on projects etc. can quickly change. Also, one mandatory property is URL, i.e. the website, via which contact information should usually be available.

Another series of iterations was necessary after an instance of TERESAH was set up and adaptations for HaS had begun. For several properties, for example, terms that were originally taken from Dublin Core were exchanged by equivalent terms from schema.org, because schema.org was specifically created to enable machine-readability and enhance findability of web content.

The AP is now finished and can be found in annex 1. Following this, the implementation of the AP within TERESAH could be completed and tests with RDFa implemented by partners in their websites can begin. As a few of those elements are bound to controlled-lists, we created a simple list of the terms that can be used, you can find the list of vocabularies in annex 2.

As a short overview here, we will go through the 22 terms of the AP, of which only four are mandatory. More information can be found in annex 1.

Name: *mandatory*, describes the name of the tool/service
Type: *mandatory*, *controlled-list*, explains if it is a tool or a service
Description: *mandatory*, provides a description of the tool/service
Url: *mandatory*, the URL where the tool can be found or the service used
Application Category: *controlled-list*, type of software application (TaDiRAH Research Activities)
Service Type: *controlled-list*, type of service being offered (HaS WP5 inkind classes)
Is Used For: *controlled-list*, describes what the tools and services can be used for (TaDiRAH Research Techniques)
Research Object: *controlled-list*, The object upon which the action is carried out (TaDiRAH Research Object)
Keyword: word that describes the tool or service or the they are used in, respectively
Standard: standards used with this tool or service
Date created: date of creation of the tool or service
Date modified: date of last modification of the tool or service
Provider: provider or operator of the tool or service
Creator: main creator/developer of the tool or service
Contributor: secondary contributor to the development of the tool or service
License: a license document applying to the tool or service
Operating System: operating systems supported
Memory Requirements: minimum memory requirements
Processor Requirements: processor architecture required
Software Requirements: external component dependency required by the application, like Java or DirectX
Browser Requirements: describes the browser requirements, for example HTML5 support
Storage Requirements: describes the storage (free space) necessary to install and run the application

4. Implementation of the Metadata Application Profile

For the implementation of the AP as part of the registry of DH tools and services, a centralised approach was considered against a decentralised one. The developed demonstrator in HaS WP8 concentrates on a decentralised approach due to different reasons. In addition to the effort of setting them up and maintaining them technically, centralised registries require much more financial and human resources to keep their content up to date. HaS explored a distributed approach where metadata about tools and services are embedded directly into the websites of their originators or projects using them. Nevertheless it is still possible to manually enter tool descriptions into the registry demonstrator TERESAH.

This is realised with the help of the W3C standard RDFa. RDFa is a well-known Resource Description Framework in the domain of the Semantic Web and allows to add structured and descriptive metadata information to (X)HTML in an easy way. Using RDFa has the benefit to augment visual information on the web with machine-readable hints by offering a set of markup attributes.[33] In comparison to microformats[34], which is also an approach to semantic markup using (X)HTML, RDFa is much more expressive. It is not limited to certain topics and allows description of nearly all domains, if adequate ontologies are used. All properties used are unambiguously identified by a URL.The barrier for the implementation in websites is quite low, as no further interface is needed for the provision of descriptive metadata and the syntax is quite simple.
This allows the demonstrator of the HaS registry for DH tools and services to realise two main parts of the intended results: the implementation of the AP in community-driven websites using RDFa and a search engine & harvester to collect the distributed information and provide them to researchers and the interested public. By developing this kind of metadata solution for the description and sharing of information on tools and services used for DH research as well as the methods supported by them, the demonstrator contributes to more openness in this domain.

5. RDFa - How to

With a finished AP, we could also approach partners as well as related websites within the fields of Digital Humanities in order to propose to them the implementation of our AP to describe their own tools and services. The description of those would then be made in RDFa[35] within their already existing websites, RDFa information would not break the frontend of their pages. When this would be done, TERESAH could then be modified in order to "harvest" those information and index them. This is where the decentralised part comes in place within our tool.

[33] https://www.w3.org/TR/rdfa-primer/
[34] http://microformats.org/wiki/Main_Page
[35] https://www.w3.org/TR/xhtml-rdfa-primer/

Explanations for webmasters

Adding RDFa to an existing website is relatively easy once you know the properties that are used by the one(s) harvesting your website, that's why we created this AP, which are the properties that can be set in HTML in order to have descriptive information easily readable by machines. Firstly, you will need to set a vocabulary at the top HTML element that encapsulates the description of your tool or service. This can be at the very top of an HTML page if the page only describes one tool / service or multiple times within the page. Once the vocab attribute has been given, it would need to be given a type attribute, our AP is set on a type of SoftwareApplication (http://schema.org/SoftwareApplication).
Therefore your top level HTML element that encapsulate the description of your item should look like the following (with @vocab and @typeof):

```
<article vocab="http://schema.org/" typeof="SoftwareApplication">
        [...]
</article>
```

Within this HTML element, you can then add all the descriptive elements you wish to have for this tool or service. All the descriptive elements you wish to use in RDFa will come from the AP, for example you wish to add a TaDiRAH Research Activities item like "Publishing"[36], you can do so by a property "applicationCategory" around your "Publishing" description:

```
<a href="http://teresah.dariah.eu/tools/by-facet/application-category/publishing"
property="applicationCategory">Publishing</a>
```

Now, by reading the AP, you will realize that in order to include a TaDiRAH Research Activities item, you will need to use http://schema.org/applicationCategory and since we are already in the vocabulary of http://schema.org, a simple property "applicationCategory" would suffice to describe that element. However, would you choose to describe the type of the item, which can only be Tool or Service (as defined by NeMO Instrument class[37]), you would need to use dc:type which is outside the scope of our current vocabulary. Therefore, a full property declaration would be needed as follow:

```
<span property="http://purl.org/dc/elements/1.1/type">Tool</span>
```

In the annex you will find a full example and its result in TERESAH.

6. Demonstrator (TERESAH)

By the end of May 2017, it was decided that HaS would reuse the tool TERESAH created by another european project (DASISH) as a basis for the demonstrator needed for the registry of tools and services.
However, reusing this tool also means redeveloping it, it needs to be reviewed in order for it to be used with the new AP as well as extending its old capabilities.

[36] http://tadirah.dariah.eu/vocab/index.php?tema=47&/publishing
[37] http://nemo.dcu.gr/index.php?p=navigate

At that time, TERESAH did only use simple text fields to describe each item, that was an important rethinking of the TERESAH backend in order to allow the use of vocabularies and calendars within TERESAH instead of only free text.

Another important development that was made on top of the original software was to add the harvesting capabilities that would allow us to retrieve information of tools and services from third party websites using the AP and be able to ingest this data into our registry.

As Supervisor or Administrator, one can also harvest webpages in order to add content to TERESAH. This page is available in the administrative section of TERESAH under the "Data Sources" drop-down menu and is labelled as "Harvester".
In the context of TERESAH, we talk about harvesting when retrieving information from a single HTML page containing RDFa data. This RDFa data can describe either one tool / service or many of those.

These harvests are done weekly on Sundays at 3am so that the data being maintained on other websites can be refreshed on TERESAH registry.
Of course, if a harvest needs to be done earlier, it can be launched manually by an Administrator or Supervisor within the Harvester's page at any point in time.

As part of reaching out to potential partners and projects, we are currently asking tool and service providers to implement the AP into their websites in order to automatically harvest and retrieve tools and services information using TERESAH.

Annex 1 - Metadata Application Profile

HaS T8.1 - Metadata Application Profile for DH research tools and services

Term / Property	From Metadata Schema / Ontology:	Explanation	Vocabulary / Range	Mandatory
Name	http://schema.org/Thing	The name of the tool/service. http://schema.org/name	Free text	Yes
Type	Dublin Core	The nature or genre of the resource. http://purl.org/dc/elements/1.1/type	NeMO instrument subclasses: - Tool - Service See "Lists of vocabularies" "Type"	Yes
Description	http://schema.org/Thing	A description of the item. http://schema.org/description	Free text	Yes
Url	http://schema.org/Property	URL of the item. http://schema.org/url	URL	Yes
Application Category	http://schema.org/SoftwareApplication	Type of software application, e.g. 'Game, Multimedia'. http://schema.org/applicationCategory	TaDiRAH Research Activities See "Lists of vocabularies" "Application Category"	No
Service Type	http://schema.org/Service	The type of service being offered https://schema.org/serviceType	HaS WP5 inkind classes See "Lists of vocabularies" "Service Type"	No
Is Used For (subject)	Dublin Core	TaDiRAH Research Techniques describes what the tools and services can be used for. dc:subject http://purl.org/dc/elements/1.1/subject	TaDiRAH Research Techniques See "Lists of vocabularies" "Is Used For"	No

Research Object (object)	http://schema.org/Thing	The object upon which the action is carried out. Used with the ontology of TaDiRAH Research Object. http://schema.org/object	TaDiRAH Research Object See "Lists of vocabularies" "Research Object"	No
Keyword	http://schema.org/Property	Keywords describing an item http://schema.org/keywords	Free text	No
Standard	http://schema.org/SoftwareApplication	Standards used in this item http://schema.org/supportingData	Free text	No
Date Created	http://schema.org/Property	The date on which the CreativeWork was created or the item was added to a DataFeed. http://schema.org/dateCreated	ISO 8601 / W3CDTF	No
Date Modified	http://schema.org/Property	The date on which the CreativeWork was most recently modified or when the item's entry was modified within a DataFeed. http://schema.org/dateModified	ISO 8601 / W3CDTF	No
Provider	http://schema.org/Service	The service provider, service operator, or service performer; the goods producer. Another party (a seller) may offer those services or goods on behalf of the provider. A provider may also serve as the seller. https://schema.org/provider	Free text	No
Creator	http://schema.org/Property	The creator/author of this CreativeWork. This is the same as the Author property for CreativeWork. http://schema.org/creator	Free text	No

Contributor	http://schema.org/Property	A secondary contributor to the CreativeWork or Event. http://schema.org/contributor	Free text	No
License	http://schema.org/Property	A license document that applies to this content, typically indicated by URL. http://schema.org/url	Free text	No
Operating System	http://schema.org/SoftwareApplication	Operating systems supported (Windows 7, OSX 10.6, Android 1.6). http://schema.org/operatingSystem	Free text	No
Memory Requirements	http://schema.org/SoftwareApplication	Minimum memory requirements. http://schema.org/memoryRequirements	Free text	No
Processor Requirements	http://schema.org/SoftwareApplication	Processor architecture required to run the application (e.g. IA64). http://schema.org/processorRequirements	Free text	No
Software Requirements	http://schema.org/SoftwareApplication	Component dependency requirements for application. This includes runtime environments and shared libraries that are not included in the application distribution package, but required to run the application (Examples: DirectX, Java or .NET runtime). http://schema.org/softwareRequirements	Free text	No
Browser Requirements	http://schema.org/WebApplication	Specifies browser requirements in human-readable text. For example, 'requires HTML5 support'. https://schema.org/browserRequirements	Free text	No
Storage Requirements	http://schema.org/SoftwareApplication	Storage requirements (free space required). http://schema.org/storageRequirements	Free text	No

Annex 2 - List of vocabularies used in the AP

HaS T8.1 - Vocabularies used in controlled-lists

Application Category	Type	Service Type	Is Used For	Research Object
Encoding	Tool	data hosting service	1_Capture	Artifacts
Gamification > Dissemination-Crowdsourcing	Service	processing service	Conversion	Bibliographic Listings
Georeferencing > Enrichment-Annotation		support service	Data Recognition	Code
Information Retrieval > Analysis-Content Analysis		access to resources	Discovering	Computers
Linked open data > Enrichment-Annotation; Dissemination-Publishing			Gathering	Curricula
Machine Learning > Analysis-Structural Analysis; Analysis-Stylistic Analysis; Analysis-Content Analysis			Imaging	Digital Humanities
Mapping			Recording	Data
Migration > Storage-Preservation			Transcription	File
Named Entity Recognition > Enrichment-Annotation; Analysis-Content Analysis			2_Creation	Images
Open Archival Information Systems > Storage-Preservation			Designing	Images (3D)
Pattern Recognition > Analysis-Relational Analysis			Programming	Infrastructure
Photography			Translation	Interaction
POS-Tagging > Analysis-Structural Analysis			Web development	Language
Preservation Metadata > Storage-Preservation			Writing	Link
Principal Component Analysis > Analysis-Stylistic Analysis			3_Enrichment	Literature
Replication > Storage-Preservation			Annotating	Manuscript

Scanning			Cleanup	Map
Searching			Editing	Metadata
Sentiment Analysis > Analysis-Content Analysis			4_Analysis	Methods
Sequence Alignment > Analysis-Relational Analysis			Content Analysis	Multimedia
Technology Preservation > Storage-Preservation			Network Analysis	Multimodal
Topic Modeling > Analysis-Content Analysis			Relational Analysis	Named Entities
Versioning > Storage-Preservation			Spatial Analysis	Persons
Web Crawling > Capture-Gathering			Structural Analysis	Projects
Bit Stream Preservation > Storage-Preservation			Stylistic Analysis	Research
Brainstorming			Visualization	Research Process
Browsing			5_Interpretation	Research Results
Cluster Analysis > Analysis-Stylistic Analysis			Contextualizing	Sheet Music
Collocation Analysis > Analysis- Structural Analysis			Modeling	Software
Concordancing > Analysis-Structural Analysis			Theorizing	Sound
Debugging			6_Storage	Standards
Distance Measures > Analysis-Stylistic Analysis			Archiving	Text
Durable Persistent Media > Storage-Preservation			Identifying	Text Bearing Objects
Emulation > Storage-Preservation			Organizing	Tools
			Preservation	Video
			7_Dissemination	VREs

			Collaboration	
			Commenting	
			Communicating	
			Crowdsourcing	
			Publishing	
			Sharing	
			o_Meta-Activities	
			Meta: Assessing	
			Meta: Community Building	
			Meta: Give Overview	
			Meta: Project Management	
			Meta: Teaching / Learning	

Annex 3 - Full HTML / RDFa example

```
<div vocab="http://schema.org/" typeof="SoftwareApplication">
 <div property="name">TERESAH Tool</div>
 <div>
   <h3>Available Data</h3>
   <dl>
     <dt>Application Category</dt>
     <dd property="applicationCategory">Transcription</dd>
     <dt>Browser Requirements</dt>
     <dd property="browserRequirements">Requires HTML5 support</dd>
     <dt>Contributor</dt>
     <dd>
       <span property="contributor">Some contributor</span>
     </dd>
     <dt>Creator</dt>
     <dd property="creator">The creator</dd>
     <dt>Date Created</dt>
     <dd property="dateCreated">2017-08-01</dd>
     <dt>Date Modified</dt>
     <dd property="dateModified">2017-08-03</dd>
     <dt>Description</dt>
     <dd property="description">This is simply a tool we needed</dd>
     <dt>Is Used For</dt>
     <dd>
       <span>Gamification</span> <!-- Value to show the user -->
      <span style="display: none;" property="http://purl.org/dc/elements/1.1/subject">Gamification >
Dissemination-Crowdsourcing</span> <!-- Value for the harvester -->
     </dd>
     <dt>Keyword</dt>
     <dd property="keywords">game</dd>
     <dt>License</dt>
     <dd>
       <a href="https://creativecommons.org/licenses/by/4.0/"
property="license">https://creativecommons.org/licenses/by/4.0/</a>
     </dd>
     <dt>Memory Requirements</dt>
     <dd property="memoryRequirements">8GB</dd>
     <dt>Operating System</dt>
     <dd property="operatingSystem">Windows 10</dd>
     <dt>Processor Requirements</dt>
     <dd property="processorRequirements">IA64</dd>
     <dt>Provider</dt>
     <dd property="provider">Provider of service</dd>
     <dt>Research Object</dt>
     <dd property="object">Images</dd>
     <dt>Service Type</dt>
     <dd property="serviceType">Processing service</dd>
     <dt>Software Requirements</dt>
     <dd property="softwareRequirements">Java 7+</dd>
     <dt>Standard</dt>
```

```
      <dd>
        <span property="supportingData">XML</span>, <span property="supportingData">JPG</span>
      </dd>
      <dt>Storage Requirements</dt>
      <dd property="storageRequirements">100GB</dd>
      <dt>Type</dt>
      <dd property="http://purl.org/dc/elements/1.1/type">Tool</dd>
      <dt>Url</dt>
      <dd>
        <a href="http://teresah.dev/" property="url">http://teresah.dev/</a>
      </dd>
    </dl>
  </div>
</div>
```

Annex 4 - Letter sent to partners and projects

Dear xxxxx,

I am writing to you on behalf of the Humanities at Scale (HaS) project (http://has.dariah.eu/). HaS aims to improve and sustain digital research in the arts and humanities by growing the DARIAH community and developing core services that facilitate better access to DARIAH contributions, support research and promote openness.

Part of the efforts of HaS takes place in the area of Open Methods. One task is to develop a metadata application profile and a registry for the description and sharing of information on DH tools and services in order to foster their dissemination and provide DH researchers with a central source of information.
For the collection of the registry contents, the focus is on a distributed approach in which the metadata are implemented directly on the websites of tools and service providers (using RDFa) and then harvested by the registry.

We are now looking for projects to support us in testing our prototype by implementing our metadata application profile on their site that we would harvest.

The application profile consists of 22 elements. Four of them (name, type, description and url) are mandatory. However, using more terms will enable you to describe your tool or service more precisely and will yield more accurate search results for future users. You can find the application profile attached
("HaS_WP8_Metadata_Application_Profile.pdf").

RDFa is used to mark up existing content on a web page to get structured, machine-readable data which also enhances their findability. To demonstrate how it works, we attached a document with a brief RDFa how-to including an example ("HaS_WP8_How_to.pdf"). More information can also be found in the RDFa primer (https://www.w3.org/TR/xhtml-rdfa-primer/).

Would it be feasible for you to implement our metadata application profile on the website(s) of your tool(s) and if so, would you consider discussing this possibility? We would really appreciate your support in this and are happy to answer any questions you might have.

Best wishes,
xxxxx