Meeting Minutes: Technical requirements workshop for the registry for semantic assets (SEMIC - A07.01)

Project:	SEMIC	Date and Time:	16/05/2024 10:00 - 12:00
Meeting Type:	Webinar	Location:	Virtual
Coordinators:	Claudio Baldassarre Nathan Ghesquière	Issue Date:	04/06/2024

Agenda of the webinar		
10:00 - 10:10	Introduction	Slides 1 - 3
10:10 - 10:20	Recap 1st workshop – Adoption Requirements	Slides 4 - 7
10:20 - 11:25	Metadata	Slides 8 - 15
11:25 - 11:45	Architecture	Slides 16 - 20
11:45 - 11:50	Summary and next steps	Slides 21 - 26

Meeting Slides
LINK

Participants		
Name	Initials	Organisation
Ana Rosa Guzman	ARG	Ministery of Territorial policy and Public Service

Arthur Schiltz	AS	PwC EU Services
Claudia Cristina Pollina	CCP	The Department for Digital Transformation
Claudio Baldassarre	CB	DIGIT
Corinna Rensch	CR	Niederrhein University of Applied Sciences
Emidio Stani	ES	PwC EU Services
Harald Pape	HP	aomation GmbH
Javier Andrés Arroyo	JAA	Ministery of Territorial policy and Public Service
Joren Verleyen	JV	PwC EU Services
Martynas Mockus	MM	Information Society Development Committee
Matthias Palmér	MP	MetaSolutions AB: private
Matteo Fortini	MF	The Department for Digital Transformation
Norman Calleja	NC	MITA (Malta Information Technology Agency)
Nathan Ghesquière	NG	PwC EU Services
Pascal Derycke	PD	Sciensano
Riitta Alkula	RA	Digital and Population Data Services Agency
Sander Van Dooren	SVD	Flemish Government

Full Meeting Minutes

Introduction Slides 1 - 3 **Speaker:** Claudio Baldassarre	Claudio Baldassarre (CB) opened the meeting and went over the agenda for the technical requirements workshop: • Recap of the first workshop and the requirements distilled from there • Metadata • Architecture • Summary and next steps During the SEMIC activities of last year, various interoperability initiatives were introduced to the Member States. By doing so, the Member States were given different options to reach a higher level of interoperability. It would take a long time to reach interoperability

	individually as all repositories have different interfaces. With the EU-wide Registry, an interface would be created for all Member States for searching models and having interactions with other Member States. With this workshop, about the technical requirements, a future pilot for connecting the national repositories and the Registry could be planned. This would improve the findability and enables models to be exposed easier.
Recap adoption requirements workshop Slides 4 - 7 **Speaker:** Nathan Ghesquière	<u>Recap first workshop</u> The previous workshop was about the adoption requirements and went over the difference between the Registry and national repositories. This was further explained by user categories that were mentioned in different use cases. <u>Captured requirements</u> During this workshop, some requirements were captured that could be used in the future pilot. These were: • Domain-specific filtering is considered important, and can be taken into the metadata; • The discoverability of models from SEMIC or the European Commission in general; • Comparability of previous model version(s) was deemed important; • The way data/models would be exposed to the Registry. This could be done by either harvesting or via an API-connection; • Prioritization of functionalities (searching and filtering over the ranking of models); • Contact information to be made available via the metadata for stimulating collaboration and co-creation; • Having the ability to declare permission, either public or private, to certain models and having the ability to have licensing conditions. When asked if some other requirements were missing, some questions were raised by the working group: Ana Rosa Guzman (ARG) asked about the Multilingualism of the model, as well as automated translations. • CB answered that this was not included, but with the search-functionality, models and labels could be looked up by connecting them to certain languages. The model should at least be in the local language, perhaps also in English. An option to document assets in more than one language should exist, as semantic experts sometimes use specific language that can be hard to understand. Matthias Palmér (MP) asked if for these requirements, the Prof Vocabulary was already looked into. This vocabulary can be found [here]().

	- CB stated that this was not yet the case, but it was noted that some additional research should be done on this. **Requirement** Models on the Registry should be multilingual, and the translation of the models should be automated. Future Pilot participants For the future pilot, SEMIC would like to thank Germany and Lithuania for already confirming their participation in the future pilot. No kick-off date for the pilot has been confirmed at this moment, but all Member States will be kept informed. The starting date of the pilot can be expected as early as from September of this year. With the pilot, a set of basic functionalities of the Registry could be tested in collaboration with the participating Member States. SEMIC therefore welcomes all Member States to participate. MP said that Sweden already has a data portal. Here, not only models but also best practices for example are shared. It contains valuable information on application profiles, ontologies, etc. MP confirms that Sweden would also be interested in participating in the future pilot, but it would depend on the approach that will be taken. - Nathan Ghesquière (NG) thanked MP for the future participation and said that it is certainly interesting for Sweden to take part in the pilot, as the Registry should be a platform that would fit everybody's needs. - MP stated that Sweden already had some interviews with different agencies and confirmed within the Prof Vocabulary, they would not use the same way of working (UML). - Riitta Alkula (RA) added that Finland is also interested but has limited resources, so their participation would depend on the scale of the pilot.
Metadata Slides 8 - 15 **Speaker:** Emidio Stani	Analysis approach Emidio Stani (ES) started by saying that research has been done on the repositories by different Member States. The researched repositories include: - DNN (Norway) - DVT (Finland) - LOV - SGF (France) - SGT (Italy) - TNO (The Netherlands) The repositories from portals that are not exclusively from Member States are LOV and TNO. Each ontology is different, while TNO is more about data spaces, LOV is the place where people could search for models and certain concepts and is more community-wide.

It also allows to find models published by certain organisations. Based on these repositories, SEMIC started to look on what metadata is used.

Adopted metadata models
Most of the repositories researched allow the use of RDF, apart from France, which uses YAML. Some parts of RDF come from the DCAT vocabulary.

Questions
- MP shared some information for the working group to look at. The information shared by MP consisted of the following additional information:
  - Minor adaptions of PROF-SE, which can be found [here](#);
  - The place on the Swedish portal where specifications could be searched, which is based on PROF-SE, and can be found [here](#).
- CB stated that some repositories do not have features that allow for searching and finding models, only features for value for metadata. Metadata should be considered a tool for designers to discover who designed certain properties. The Registry would detail what is in the model and the importance of having access to individual classes, concepts, properties, etc.
- RA said that Finland has a new version of the DVT/Data model tool. A beta version has already been launched.

Requirement
Having access to individual classes, concepts and properties could be important.

Common & uncommon metadata
ES started by defining the common metadata, which were 'Name' and 'Description'. These are metadata found in every repository analysed. With some repositories, these metadata were also translated. ES stated that, as mentioned before, language and multilingualism is something good to leverage.

Less common metadata refers to metadata that were not found in every repository but were still found in some of the repositories. These are among others 'Version_info', 'Publisher', 'Used by', which gives information for what and by whom the data model is used and 'Contact point', which allows future collaborations.

The last category of metadata is the suggested metadata. These were not found in the analysed repositories, but they could be added to the Registry. The suggested metadata is 'Licence'. It was found in

some portals but was left up to the people providing the models and metadata.

Questions

MP mentioned that he did an investigation where he compared several initiatives. They can be found [here](#) (in Swedish).

Sander Van Dooren (SVD) said that Joinup used to be ADMS-AP based, but he is not sure if this is still the case. He also said that Flanders has an endpoint which contains the actual models. There is no metadata there at this point unfortunately. The metadata is stored elsewhere in machine-readable format.

Claudia Cristina Pollina (CCP) stated that there is no licence for every asset in Italy, but a general description can be found in a legal note.
- CB iterated that add a description of each model in a legal note is something that should be done.

CB then asked if members of the working group thought of other metadata that could be leveraged and could be interesting for the Registry.
- MP mentioned that, in contrary to the Prof Vocabulary, there is no distinction between the specification of a model and specification parts, such as classes and properties.
- ES said that there are application profiles that can describe the same metadata. In the Norwegian repository for example, there is some common metadata among the artefacts and it has a hierarchical concept (information model being on top). In other repositories, the metadata is shared without hierarchical concept.
- MP reiterated that it is important to connect specifications to datasets, so the dataset is conform with the metadata. It could be linked to a PDF or specific document, or even an UML that has all classes defined. This is partly done by Prof, which combines different resources.

After that, Pascal Derycke (PD) asked if any Controlled Vocabulary is in use in theme.
- ES answered that some models do, but it depends on the classifications.
- NG added that this is a decision that will be made once the pilot is in process.
- SVD said that the Publications Office has a data theme controlled vocabulary, which can be found [here](#).
- MP added that a data themed vocabulary has been added in PROF-SE.
- SVD mentioned that this vocabulary is still quite minimalistic.
- MP stated that this could be complemented with concepts from a vocabulary like EuroVoc for example. Another option could be to complement it with keywords.

Requirement
The Registry could provide the ability to connect model specification to datasets. An example could be the link between data/europa.eu and the models behind the datasets.

To gather information, NG introduced several questions about the metadata components.

Slido question 1
The first question was about the uncommon metadata and which ones should be mandatory for the Registry. The results from this question resulted in some prioritizations. 'Name' and 'URI identifier' are voted by every participant to be mandatory. The next four most metadata components deemed important are:
- 'Description'
- 'Languages'
- 'Publisher'
- 'Contact point'

Other metadata were partially voted, but 'Contributor' only had 1 vote and 'Used by' did not get any votes.
- MP said that "Used by" should be calculated by finding datasets that point to it via dcterms:conformTo. He therefore thought that it should not be an explicit attribute.
- ARG asked if proprietary models would be accepted.

CB stated that the participants that did not vote might not have the proper knowledge on all the metadata presented. He also stated that the priority created by the participants on which metadata is more important to be mandatory could be neglected. NG said that going over the options created by the working group could still be useful.

An overview of the answers given on which metadata should be mandatory can be found below:

Which of the following metadata components should be mandatory?

Multiple Choice Poll   6 votes   6 participants

Name - 6 votes

100%

URI Identifier - 6 votes

100%

Description - 5 votes

83%

Languages - 5 votes

83%

Publisher - 5 votes

83%

Contact point - 5 votes

83%

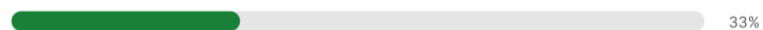Publication date - 4 votes
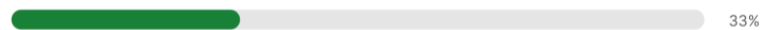
67%

Modified date - 4 votes

67%

Theme - 4 votes
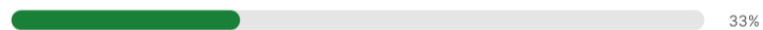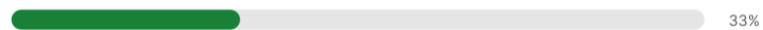
67%

Version_info - 3 votes
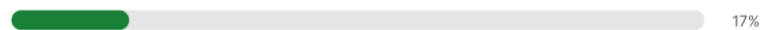
50%

Creation date - 2 votes

33%

Main concepts - 2 votes

33%

Status - 2 votes

33%

Access rights - 2 votes

33%

Contributor - 1 vote

17%

Used by - 0 votes

0%
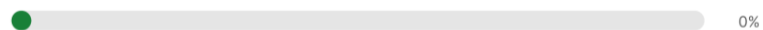
Slido question 2
NG explained that the second question was about the metadata that should be optional. MP then asked if there will be another question that will ask if the metadata components should be recommended. NG answered that in this exercise, recommended and optional can

be considered the same, namely something that is not mandatory but would be nice to have. MP followed up by asking if it should be considered to not vote for the components that people previously voted as mandatory. NG responded that it is not necessary to vote again for the mandatory ones.

While the working group was voting, CB explained that he would rather have all metadata presented today used in the registry, and some being mandatory, than restrict the number of metadata.

- MP asked if the metadata would be for the entire model or for specific classes only.
- ES responded by saying that it would be for the entire model.
- MP then asked if the model would thus be a simple indicator for potential reuse or if it is more complicated than this.
- CB answered that this could be seen as a level of access; which usage could be investigated if needed.
- CB went over the results, saying that if a certain metadata component has for example 80% preference to be optional, which was the case for the metadata component access rights, it is important to consider for the Registry.
- Norman Calleja (NC) said that in Malta, there usually is a default license on the portal since the data is visible. It could be a possibility however to indicate some restrictions in the use of certain datasets.
- ARG asked if someone had an example of a sensitive model where access rights were deemed important. The Registry would be for the public sector, where data models are publicly available.
- MP said that certain models are contained by public entities, but models in the healthcare domain are not per se public. Some models had to be bought to make available and public to the community.
- ARG stated that this was not under the concept of access rights but should be considered more under license. If the registry decides to provide private models, then having access rights as metadata would become a different discussion.
- CB said that if the registry provided features of search, there is a difference between how optional and mandatory are perceived. Mandatory would mean the metadata is required and optional would be more to fill in the metadata for the model in a FAIR-sense. He continued that within the scope of the Registry, interoperability and discoverability should be promoted. To enable this, the definition about what metadata should be, should be worked on.
- MP said that mandatory could be considered what is necessary for the Registry to work and optional should be what is preferrable but not necessary. It has the same definition in the Swedish data models
- ARG said the working group is being too cautious. There are good metadata models for reference available but vocabularies that describe these assets are not being used. Thinking in public

	services in general does not cover all sectorial technicalities and the focus of the Registry should be a general point of view. ARG asked what the exact scope of the Registry is.
	• CB reassured that the scope is still the public sector. There is a possibility for the scope to be widened, but this would create problems with the initial intentions for the creation of the Registry. Access rights should be defined and explained why it would be taken into the metadata. What would be exposed on the Registry is also publicly available on the national repositories.
	Requirement
	• Access rights should be defined so the definition is clear. It should also be explained why it is taken into the metadata.
	The answers given for this question can be found in the visual below

Which of the following metadata components should be optional?
Multiple Choice Poll ☑ 5 votes 👥 5 participants

Acces rights - 4 votes
████████████████████░░░░░░ 80%

Version_info - 3 votes
███████████████░░░░░░░░░░░ 60%

Creation date - 3 votes
███████████████░░░░░░░░░░░ 60%

Publication date - 3 votes
███████████████░░░░░░░░░░░ 60%

Modified date - 3 votes
███████████████░░░░░░░░░░░ 60%

Main concepts - 3 votes
███████████████░░░░░░░░░░░ 60%

Status - 3 votes
███████████████░░░░░░░░░░░ 60%

Contact point - 3 votes
███████████████░░░░░░░░░░░ 60%

Contributor - 2 votes
██████████░░░░░░░░░░░░░░░░ 40%
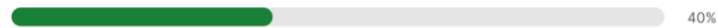
Publisher - 2 votes
██████████░░░░░░░░░░░░░░░░ 40%
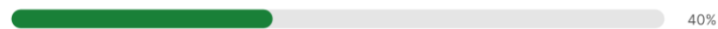
Theme - 2 votes
██████████░░░░░░░░░░░░░░░░ 40%
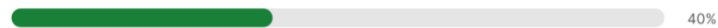
Used by - 2 votes
██████████░░░░░░░░░░░░░░░░ 40%

Description - 1 vote
█████░░░░░░░░░░░░░░░░░░░░░ 20%
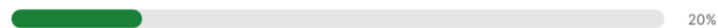
Name - 0 votes
░░░░░░░░░░░░░░░░░░░░░░░░░░ 0%
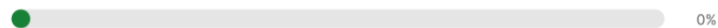
Languages - 0 votes
░░░░░░░░░░░░░░░░░░░░░░░░░░ 0%

URI Identifier - 0 votes
░░░░░░░░░░░░░░░░░░░░░░░░░░ 0%
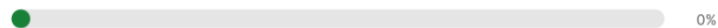
Slido question 3
The third and last question was an open question to the working group. It was an open question about potential metadata components that might have been forgotten in SEMIC's analysis.
CB reiterated that with the future pilot, the bar will be lifted from a technical level to a conceptual level.

The questions asked by the working group were:

Regarding the third and fourth question, CB answered that because the working group is comprehensive, it gives the working group the ability to talk about everything in a certain way. CB stated that he likes to be more expressive, saying that it is better to define what is mandatory and recommended and leave all the metadata open. The metadata that is necessary to talk about a certain model should be defined. The indication should thus be on what is mandatory and what is optional, while not restricting anything.

ES mentioned that there is the concept of distinction in DCAT that talks about access rights. One is access rights while the other is license, so it should be investigated more.

- Martynas Mockus (MM) believed that the territorial classification might be missing. A specification could be made whether the semantic asset pertains to the EU-area, a specific country, a region, etc.
- MP agreed that the origin of a model was indeed missing.
- NC responded that he understood it as jurisdiction
- MM said that jurisdiction is coming from the Justice data space.
- ARG said that it is unclear whether a model (not the data modelled by it) belongs to a jurisdiction. When describing a data

	space, it makes sense to have jurisdiction, but not when describing a model.
	• SVD said that the territorial classification is covered by ADMS-AP. There is however a difficult understanding of this property, namely where the model originated, where the model is compatible with the legal context or where the model would like to be used. Therefore, the territorial classification should be clearly defined.
	• CB responded by saying that hard to understand how a model can be applicable to a certain jurisdiction. He wondered how a model could be bounded to different laws. If a model is 'valid for a certain territory', it is more information for the person/people who created the model. CB also agreed with ARG, saying that it makes no sense to restrict a public model to a jurisdiction. Certainly not when the intention of this jurisdiction is the prevention of reuse.
	• MP mentioned that when someone is maintaining a model, various aspects should be taken into account. The model could be on different levels (national or topical level for example). DCAT-AP, which is on metadata level, should be on national and European level. This way, the properties should be translated and conceptualised. This information would not per se be in the model but could also be in an adjacent document. MP continued saying he does not like the term models and prefers the term specification part. If you have a specification, it does make sense to have this term. Shacl shapes could be seen as metadata and with this, when the model is in the system, it can be seen as different semantic assets that are connected with each other.
	• CB answered that it is already conceptualised that the Registry should not contain models. The models would remain in their respective repositories. This way, it could be a good tool to discover the values of reuse of the same concepts in different vocabularies. You could have a network of connections with the different national repositories, giving these repositories instruments on to count how many times a certain model or concepts or used. This could be useful to get an idea on the potential visibility of a model.
	• MP continued by giving an example. He stated that with the Core public event vocabulary (CPEV), it would be nice to see if certain properties/classes are reused in similar models. For this to be accomplished, the model would need to be downloaded to be contextualised.
	• CB reiterated that the ownership of the models should remain with the Member States. If other Member States would have the opportunity to access these models, they could navigate to the most adopted model.
	Requirement

	• A distinction between access rights and licenses should be made and clarified. • There should be a clarification on which level (national, local, ...) a model within a Member State is stored. • Ownership of the models should remain with the respective Member States.
Architecture Slides 16 - 20 **Speaker:** Emidio Stani	Components ES said that the components presented are from the point of view from the repositories that were researched. It was found that, for these repositories: • The internal database within the repositories was mostly used for different purposes; • Most of the repositories use a Sparql endpoint. This endpoint was public in most repositories, but private in some; • Most of the repositories also provided a REST API; • Lastly, some of the repositories researched provide a validation service. For harvesting or accessing the metadata, based on these repositories, the preferred way is via Sparql endpoints (RDF). Upstream & downstream scenarios To connect the Registry and the repositories, two options could be used. It could either be via a Sparql endpoint (RDF) or via a REST API. To determine the traffic with both, authentication keys should be in place. This could for example be an API gateway that needs to be passed before being able to query the Sparql endpoint. The traffic in a gateway should be monitored so data can be retrieved when there is less traffic than usual. Certain things need to be solved in order for this to work, because the repositories need to know how they can access the Registry. CB added that Member States must be willing, and have the capacity for this to work. Creating access to the Registry is a technical effort and comes down to transferring the message of potential and opportunities to be part of this. *Questions* SVD and RA asked if the registry could also be exposed as LDES. Belgium and Sweden are working on integrating them • ES answered that the Sparql endpoint is preferred, as starting from something that exists is essential. SVD also mentioned that a flat file would be the easiest way and manageable for everyone. This way, a file could be dumped where someone else could pick it up. If a Member State uses a Sparql endpoint, one simple line could be created to access this file. • CB said that he will not exclude this way of working, even if it meant for the Registry to be a data space. It could overlap with

	what is done for the Registry, and if after investigation it does, a guidance could be created on how to access the national repositories. • SVD mentioned that it does not depend on where the model is located, but rather that the components in data spaces are not entirely the same. SVD continued by saying that he felt uneasy about the way semantic assets remain semantic assets. Joinup used to have harvesters, but the information that is available should be used. • CB said that it would be clearer when the ends are put together and when Member States participate in the future pilot. The quickest way to progress is to reuse existing stuff, find limitations there and tackle them.
Summary & next steps Slides 21 - 26 **Speaker:** Claudio Baldassarre	<u>Roadmap</u> CB stated that this was the second workshop and the third session of the working group. He reiterated that this working group is not meant for making definite decisions, but rather to make the creation of the EU-wide Registry a community effort. This could only be done by having healthy discussions, as was the case today. <u>GitHub</u> CB continued, saying feedback can always be provided via issues on the GitHub dedicated to the Registry (issues can be found [here](#)) or via mail. <u>Closing</u> To close this meeting, CB thanked everybody attending for being active and engaging in discussions. Any form of interactions, whether it is within this working group or outside of it is more than welcome. The next meeting will be the last meeting of this working group. It will consist of a shorter closing meeting, where people from outside this working group will be invited for to give potential additional insights. The practicalities of this meeting will be communicated on a later date.