SEMAF: A Proposal for a Flexible Semantic Mapping Framework

Version: 1.0, March 2021

Authors

Name	Affiliation	ORCID
Broeder, Daan	CLARIN ERIC	0000-0002-8446-3410
Budroni, Paolo	TU Wien	0000-0001-7490-5716
Degl'Innocenti, Emiliano	CNR-OVI, ERIHS	0000-0002-3839-9024
Le Franc, Yann	e-Science Data Factory	0000-0003-4631-418X
Hugo, Wim	e-Science Data Factory, KNAW/DANS	0000-0002-0255-5101
Jeffery, Keith	Keith G Jeffery Consultants, EPOS-ERIC and British Geological Survey	0000-0003-4053-7825
Weiland, Claus	DiSSCo, Senckenberg	0000-0003-0351-6523
Wittenburg, Peter	Max Planck Computing and Data Facility	0000-0003-3538-0106
Zwolf, Carlo Maria	LERMA, Observatoire de Paris, PSL Research University, CNRS	0000-0002-5762-6747

This report was created under the responsibility of CLARIN ERIC (clarin@clarin.eu). For comments and questions please contact Daan Broeder d.g.broeder@uu.nl .

Acknowledgements

This report has been produced with the support of EOSCSecretariat.eu which has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, Grant Agreement number 831644.

Contents

Executive Summary

We recognise an enormous increase of differing semantic spaces defined by a variety of semantic artefacts ranging from large ontologies to lists of terms. These are being created by individual researchers or research communities, based on theories, to capture relevant phenomena or to serve some pragmatic needs. These semantic spaces serve relevant functions in some research communities and cannot be changed easily due to pragmatic needs (such as preservation of vocabularies as used at any one time) and as they may be utilised in existing semantic crosswalks. Nonetheless, these community semantic spaces and toolsets evolve with time, hence curation and provenance are important.

We also recognise an increasing need to allow the joint use of data coming from different domains applying such semantic artefacts. This requires a mapping between concepts emerging from the different approaches. Such mappings are already very common within larger research communities such as health where, for example, cultural and language differences need to be bridged, but seem to be also relevant for an increasing wish to integrate data from different disciplines such as, for example, ethnological studies with climate change data. The potential number of such crosswalks is very large and thus the creation of an overall ontology is not practically possible or theoretically feasible.

The solution is to establish a flexible semantic framework that is driven by pragmatic considerations, can link up to different kinds of semantic artefacts, can be used by all researchers or research groups and that follows FAIR principles. Many projects are already doing semantic mapping especially in the field of metadata, but in some cases also in the field of data. Yet, with very few exceptions, these mappings are hidden in non-FAIR data structures or in software created by domain experts. Thus, these mappings are not explicit, and cannot be shared.

25 interviews with key researchers and data managers from different fields confirmed more formally what could be observed from earlier studies and discussions: .such a flexible semantic mapping framework is a necessary addition to the existing practices since it would:

- make yet hidden mapping schemes explicit, shareable and reusable;
- enable researchers to make crosswalks easily using (semi-)automatic alignment tools without waiting on semantic experts;
- assist researchers to register and publish schemes.

Such a mapping framework - being developed under the guidance and oversight of the EOSC process guaranteeing a proper and shared specification of a mapping registry - will be necessary for increasing interoperability. It is not a trivial task since a set of important requirements need to be fulfilled, and some challenges need to be addressed, as expressed in the interviews. We suggest an inclusive and modular approach that will allow using the registry and mapping description functionality separately from the actual mapping implementation and mapping operationalization and tooling, since these have separate governance and development dynamics. Separation will also allow pragmatic and selective integration of components.

To realise SEMAF, we suggest in addition to a short term design study, a 3 year project to EOSC to establish a registry and supporting tools - with appropriate outreach and training - to achieve the needed uptake.

Introduction

Overview

This document is the final report of the SEMAF project (a study funded by the co-creation programme from the EOSC secretariat #14) to develop a proposal for a flexible semantic mapping framework.

The background and motivation for this work comes from discussions between research infrastructure experts - especially in the context of the Research Data Alliance[1] (RDA) - who detected a gap w.r.t. a pragmatic approach to semantic interoperability and easy-to-use facilities in the current data infrastructure landscape. The EOSC secretariat's call for co-creation projects offered a chance to address this through proposing a project studying requirements and interest in a framework offering a pragmatic approach to semantic mappings as a first step towards such a framework. The initial group requested CLARIN ERIC to submit the proposal for studying a Semantic Mapping Framework (SEMAF) to the secretariat. The initial group was subsequently extended with other experts into a SEMAF task-force, whereof the members are all authors of this report.

Goal, Process

The main objective of this project is to specify a flexible framework to create, document and publish mappings and cross-walks linking different semantic artefacts within a particular scientific community, as well as across scientific domains.

The work plan drawn up by the task-force specified the following steps and phases:

- Extended discussions amongst the SEMAF task-force members, supplemented by existing literature as well as new material produced in the context of the FAIR and EOSC discussions[2] to establish a common basis for the goals and scope of the project work.
- The creation of a SEMAF mission statement document, highlighting the motivations for the SEMAF project, and providing a basis to discuss and inform experts and informants outside the immediate SEMAF task-force of the topics at hand.
- A list of community data experts and researchers was created to guide a series of interviews. These served to expand our knowledge w.r.t. semantic interoperability and

[1] https://www.rd-alliance.org
[2] See the bibliography

mapping challenges faced by and implementations used by the different research communities, and provided an opportunity to discuss the proposed SEMAF solutions on the basis of the SEMAF mission statement.

- The analysis of the interviews was used, together with the results of the task-force's initial discussions, to improve and refine the task-force's initial set of requirements and plans.
- Inclusion of a proposal for follow-up activities for actual implementation of the SEMAF framework that is based on the information provided by our community expert informants and their knowledge of existing components and services that can be used as examples or be integrated into a solution.
- Soliciting feedback from the interviewed experts on the basis of a first draft of the report.

There are limited plans in place for follow-up SEMAF activities related to uptake of our results, and to further discussions with research communities. These activities will take place outside the context of the funded co-creation project, but will be aligned with other EOSC projects and the RDA working groups. Such activities should be intensified with follow-up funding.

Vision, Scope and Limitations

SEMAF's vision is to define a conceptual model for managing the mappings between semantic artefacts (typically vocabularies, lexicons, thesauri, ontologies). We wish to have a model that is implementation agnostic, acknowledging that there are many possible logical models (using different kinds of technology) and physical implementations (using specific technologies). Although triples are commonly used today to represent semantic artefacts, richer representations are emerging. SEMAF should not constrain evolution to improved approaches and technologies.

Although the SEMAF project (for now) is limited to a study for a flexible mapping framework only, based on the task-force expertise and discussion with community experts, we are confident that its proposals for follow-up implementation are realistic and will meet the requirements of a broad set of communities.

Organization of this report

The structure of the final report partially represents the different phases of the project process, and while we in general tried to avoid it, for aspects of legibility of the chapters repetition of some information was thought acceptable. Chapter "Community experts interview analysis" represents our view on the status of the communities and the views of the interviewed experts, this is not always aligned with our recommendations and proposals for the SEMAF framework that we present in Chapters "SEMAF Requirements", "Proposed Architectural Outline" and "SEMAF Proposed Next Steps". We appreciate that the terminology used to describe semantic interoperability and data infrastructure in general can differ depending on terminology used in this document.

Problem Statement

In the current data landscape we find that the data creation process in different research communities is mostly determined by established practices and not by considerations with respect to interoperability and sharing, although FAIR initiatives have made communities much more conscious of these sharing aspects. The basic needs of specific research communities, or by data providers augmenting their services, remains more influential.

This may have resulted in efficient data processing and procedures within the boundaries of such communities, but inherently pose problems w.r.t. large scale FAIR data sharing[3], open verifiable science, and cost-efficiency. Within individual research data management silos, there is an increased need for integration of data originating from other research communities - either for purposes of comparing data sets, or because research data facilities are given wider responsibilities. Challenges with respect to data interoperability are often separated into “syntactic' and schematic' vs. “semantic” interoperability problems, and it is especially the semantic interoperability which is difficult to solve as well as validate in a manageable and cost-effective way.

Some examples of the use of mappings are provided below.

Semantic mapping is often guided by pragmatic considerations motivated by the kind of research at that very moment. Linguistic research is often cross-language, research which creates problems due to the fact that the languages being compared belong to different language families. For example, some languages do not know the concept of "adverbs". Supposing the concrete research task would include to ignore the difference between "verbs" and "adverbs" known in many languages, the researcher may create a mapping assertion "adverb is_a verb". Now the researcher could do his statistical analysis, conscious of what he is doing and how to interpret the results.

[3] Wilkinson et al. (15 March 2016). "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*. **3**: 160018. doi:10.1038/sdata.2016.18.

Another example (Table 1) shows pairwise mappings of ontologies from Biodiversity and Earth System Sciences.

Entity 1	Entity 2
tectonic movement(ENVO:01001093)	Continental drift (SWEETPhenGeolTectonic:ContinentalDrift)
river bank (ENVO:00000143)	Riparian zone (SWEETRealmLandCoastal:RiparianZone)
marine benthic biome (ENVO:01000024)	Benthic zone (SWEETRealmOcean:BenthicZone)
leaf alternate placement(FLOPO:0001032	Phyllotaxy (TO:0006014)
rhizome mass (FLOPO:0003190)	Rhizome dry weight (TO:0000556)
whole plant lifestyle (FLOPO:0980070)	Life cycle habit (TO:0002725)

Table 1: Example of pairwise mappings of ontologies from Biodiversity (Flora Phenotype Ontology/FLOPO and Plant Trait Ontology/TO) and Earth System Sciences (Environment Ontology/ENVO and Semantic Web for Earth and Environment Technology Ontology/SWEET). Mappings were created for the Biodiversity and Ecology track (biodiv) of the Ontology Alignment Evaluation Initiative (OAEI, [39]).

Such problems focus on interoperable metadata and alignment of observable phenomena and simulation measurements, although accents differ per discipline. Additionally, the data semantics often remain implicit by being described in documentation that is difficult to access, especially for external users, or conveyed by oral transmission between researchers.

Exhaustive solutions for semantic interoperability, such as matching all concepts and terms to a single all-encompassing ontology, may be interesting from a theoretical point of view, but in practice creating and maintaining such ontologies is quite expensive, are difficult to agree on, and are often an overkill for achieving limited project goals w.r.t. semantic interoperability. The alternative, of providing crosswalks between every pair of semantic spaces leads to *n*(n-1)* crosswalks - clearly unsustainable when it would be required for the full semantic space. A 'half-way-house' is for each domain with multiple semantic spaces to choose a rich canonical formalism and convert each semantic space to this one, resulting in *n* crosswalks per domain which is more manageable. Cross-domain mappings required are then limited to the number of domains. Some information loss can result when mapping to a less rich formalism ', and conversely, using a very rich formalism may become a superset of all the other semantic spaces and be unwieldy.

One strongly advocated strategy for semantic interoperability is the use of Linked Data and the Semantic Web standards[4] such as RDF and OWL - provided by W3C - to describe and provide data semantics and relations and use inference and SPARQL queries to translate between different semantic domains. This is, of course, an excellent approach, but not all required relations are available or can be automatically inferred from existing openly available mapping relations. Providing the necessary mapping relations often remains a separate and specific effort. More recent research moves towards property graph or knowledge graph

[4] https://www.w3.org/standards/semanticweb/

technologies are intended to overcome the representational difficulties with RDF and triples as are temporal-RDF and OWL to cope with semantic drift.

Because of the challenges related to the all-encompassing solution strategies, the semantic interoperability challenges for individual projects and disciplines are often solved in pragmatic but limited ways i.e. only solve the problem at hand, and do not address the interoperability problem at large. Rather than advocating and imposing a general ontology, there is usually a 'translation' between those (parts of) metadata descriptions and observation measurements that are actually required for a specific goal. Such translations or mappings[5] may take the form of XSLT, RDF or software implementations, but also documentation instructing data managers how to convert records between different schemes. Such pragmatic solutions often remain invisible to the larger community, are not easily shareable, and certainly don't meet the FAIR requirements[6] as for proposed for semantic artefacts[7] [6]. Ongoing initiatives such as the RDA metadata catalogue[8], that lists a number of existing metadata schema conversions, provides examples of all of these approaches.

To improve this situation our proposal is to work towards a framework supporting such pragmatic solutions and fostering their registration, proper description, formalization, and sharing. As we note from our interactions with different communities, just finding and sharing properly described solutions including their provenance would already be very useful. Next to such registries one or more supported mapping technologies, that would permit easy sharing and reuse of existing mapping solutions would be needed to give the infrastructure direct impact.

Since the interoperability solutions that are currently used and available from the communities are very diverse and ranging from complex tools that generate special conversion logic, to human readable descriptions, such an interoperability solution registry and infrastructure should not make unnecessary exclusive choices w.r.t. technologies used. In principle all mapping implementation technologies currently used should be supported if possible, unless it is contrary to the FAIR and Open Science principles[9] [9] (e.g. no proprietary software etc.). Some technologies will offer better operationalization options than others, but remaining (for now) technology agnostic will maximise the integration of existing mapping components and services.

[5] In this document we prefer to use the word mapping
[6] https://www.force11.org/group/fairgroup/fairprinciples
[7] FAIRsFAIR project deliverable D2.2 "FAIR Semantics: First recommendations" from doi:10.5281/zenodo.3707985 (2020)
[8] https://rdamsc.bath.ac.uk
[9] FOSTER Consortium (26 November 2018). "What is Open Science?". *Zenodo*. doi:10.5281/zenodo.2629946. Retrieved 13 August 2020

Community Experts: Interview Analysis

In total we conducted 25 interviews distributed over biology, biomedical sciences, environmental sciences, climate science, natural sciences, humanities, social science, cultural heritage, and generic data management & science. Our analysis of the interviews showed that the diversity of approaches and solutions does not allow one to obtain quantitative answers. See table 2 for a list of experts that contributed.

Julian Richards	Director of the Archaeology Data Service	Humanities, Archaeology
Christian Ohmann, Steve Canham	ECRIN	Biomedical sciences
Ingemar Häggström, Carl-Fredrik Enell	EISCAT Scientific Association	Natural Science, Environmental Science
Alexandra Kokkinaki	BODC	Environmental Science, Oceanography
John Watkins	UK Centre for Ecology and Hydrology	Environmental Science, ecology, hydrology
Menzo Windhouwer	KNAW/HuC, CLARIN ERIC	Humanities, linguistics
Wolfgang Schmidle	DAI, Data scientist	Humanities, Archaeology
Johan Fihn Marberg	SND, CTO	General research data management, social sciences
Matej Durco	OEAW	Humanities
Dieter van Uytvanck	CLARIN ERIC, CTO	Humanities, linguistics
Baptiste Cecconi	OBSPM	Planetary science
Mathias Dillen	DiSSCo, MBG	Biodiversity, Environmental Science
David Fichtmüller	DiSSCo, BGBM	Biodiversity, Environmental Science
Carsten Thiel	CESSDA ERIC, CTO	Social Sciences
Herve L'Hours	UKDA	Social Sciences
Tobias Gradl	University of Bamberg	Research Data Management, Digital Humanities
Daniel Heydebreck, Anna-Lena Flügel, Claudia Martens	DKRZ	Climate science, general research data management

Ilaria Rosati, Nicola Fiore, Lucia Vaira, Pierfrancesco Tommasino	LifeWatch-ERIC, National Research Council of Italy	Biodiversity, Environmental Science
Dr. Helen Parkinson (Head of Molecular Archival Resources)	EMBL-EBI	Biomedical sciences, Bioinformatics
Margareta Hellström, Researcher Oleg Mirzov, System Architect	ICOS Carbon Portal and Dept of Physical Geography and Ecosystem Science, Lund University	Environmental Science
Federica Spinelli, Alessia Spadi	RESTORE project, OVI, National Research Council of Italy	Humanities
Claudia Caliri	ISPC, National Research Council of Italy	Cultural heritage science
Carsten Baldauf	Nomad Project, FHI	Material Sciences
Lara Ferrighi	Data Manager, Norwegian Meteorological Institute	Meteorology
Chris Schubert	Head CCCA data center	Climate sciences

Table 1: Experts interviewed and their affiliations

Communities were additionally represented via the SEMAF task-force members eg. Keith Jeffery provided a summary of his discussions with experts in EPOS (Geoscience) which apply generally to the ENVRI cluster.

Overview

Having analysed the 25 interview reports, we can state that the major impressions from having deeply studied 75 different research infrastructure reports earlier this year is mainly confirmed. We refer to the insights of K. Jeffery et al [11] via two citations:

(1) Researchers in almost all fields are using sensors that create increasing amounts of data and are increasingly willing to share this raw data. Yet there is too little awareness that the context of the experiments (lab notebook, sample preparation techniques, sensor configurations, etc) also needs to be FAIR and shared to allow other researchers to truly understand the data, assess its usefulness (relevance, quality) for their purpose and for reproducibility, for example. A cultural change is required to convince researchers to offer this kind of contextual knowledge which is still seen as private information.

(2) Most data labs are still focusing on their immediate needs emerging from discipline specific research questions and are hardly thinking in terms of usability beyond their own narrow

boundaries. Fostering interdisciplinary research requires some altruism since sufficient contextual information needs to be associated with data.

These impressions are confirmed by our interviews and have as a consequence that semantic explicitness of concepts, vocabularies and relations so that others including machines can use them is still in its infancy. Lots of different methods, styles, formats and organisations are being used so that FAIRness is a distant goal. The interviews indicate that an increasing number of researchers are aware of the needs of Open Science, i.e. making digital artefacts available beyond their immediate group interests, and the gaps that need to be overcome. But we can also see that many lack human resources and hesitate to invest much effort in this. This is aggravated by a combination of the uncertainties about technology choices, too many projects being launched resulting in a spectrum of suggestions of how to address challenges, a shortage of expensive experts, and especially a huge lack of systematic tool support.

After many decades of semantic technology development and the accompanying somewhat inflated expectations from the Semantic Web movement, we still see only a limited impact of this in current data practices. Limited but notable exceptions are the explicit provisioning by many of vocabulary resources in W3C semantic standard formats, the occasional use of vocabulary matching, and use of a specific ontology to ground all metadata used in a repository.

There is also broad realization in specific domains such as the Humanities that, in general, semantic mapping can be complex i.e. going beyond the equivalence of two elements or values in a metadata schema, and needing deeper context and complex logic such as for instance in the case where the parents and siblings of an element in an hierarchical (metadata) schema influence its semantics and thus any applicable mapping rule.

One should also reduce unrealistic expectations with respect to semantic processing. In general, we share L. Floridi's [8] remark that our current computational approaches (machines) can only be seen as "syntactic engines" and not "semantic engines". His argument is that current approaches of "semantic processing" do all kinds of more or less simple structural operations (term matching, graph matching, simple procedures based on very limited relation types, etc.). Lab practices also in those cases where new and promising "semantic" tools are being used are far away from the ideals of the Semantic Web as described by Berners-Lee [5].

Observations from Community Practices

We distinguish between semantic mapping in the metadata domain and in the domain of content encoding, observation measurement variables, annotations etc. This is done for practical reasons - in principle the fundamentals for both mapping metadata and content are the same. While metadata mapping is often already practiced for instance where discipline-wide research infrastructures with a community-wide mandate need to provide consolidated metadata catalogues, semantic mapping of content is done mainly in individual research projects or software silo level. Therefore, we will first focus on metadata and then on content observations.

Metadata Observations

In many communities research infrastructure projects have been started some years ago and in almost all cases the development of metadata catalogues was one of the pillars to achieve better visibility of data. This implied that metadata from different sources or repositories needed to be harvested, curated, mapped and indexed. Inter-converting metadata is where most progress has been achieved - firstly because of its necessity (and accompanying available resources) for the consolidated metadata catalogue effort, and secondly, because of the relatively low number of concepts involved with high level metadata descriptions[10].

Metadata conversions using mapping are carried out mostly by centres applying different technologies without, however, making the mappings explicit, shareable, and changeable. Mappings are often required from centres to satisfy the varying needs of metadata harvesters with respect to semantic spaces and formats (DataCite, DC, RDF triples, CIDOC, DCAT, CMDI, DDI, INSPIRE, CERIF, etc.). Mappings are either encoded in tables augmenting the catalogue software stack, directly included as logic in the software, partly since also procedural (conditional) criteria need to be applied or now, in a few cases as RDF assertions that can be integrated into SPARQL endpoint databases. The SKOS formalism is being used in some cases to map different labels to common concepts. In some cases simple XSLT transformations are being used to implement the mappings. Special thesauri with synonyms can provide mappings for query expansion. In one case the CERIF [12] framework is being used to create and maintain the mappings.

As indicated, in some communities and cases, metadata mapping cannot be done without using the deeper context in which the metadata schema elements occur. To handle this properly, complex logic is required, supporting conditional constructs.

One aspect of these central harvester approaches is that the centres feel obliged to provide solutions for the whole community requiring a process of inspection, consultation and mapping adaptations for specific metadata variants which in some cases is seen as limiting opportunities.

Content Observations

Common solutions for content mapping hardly exist, although the need especially in the life sciences (medicine, bioinformatics, biodiversity, etc.) to enable crosswalks between different concepts and vocabularies - emerging partly from different languages and cultures - is pressing. Many ontologies of different types have emerged over the last decades and they are being maintained and used in a variety of disciplines, but in this report we will not discuss the use of ontologies in data practices, but rather the methods of supporting transformations by semantic mapping. It should be noted, however, that proper semantic mapping without the existence of explicit ontologies is not possible.

[10] Note that when creating more precise metadata descriptions the situation changes and for instance the different domain and format-specific vocabularies necessary can be huge and alignment does not scale

Silo-based Semantic Mapping is being Done!

- Semantic mapping[11] is being done broadly in many disciplines, but it is mainly done in isolation either at individual, project or department level, i.e.
    - The semantics of the schema elements and values are often not well described nor explicit, but exist in the heads and minds of the data management experts.
    - Data providers do not see the need yet to invest efforts in empowering transformations between semantic domains outside of their ingroup, since the need for this until now is low and funding is not available; therefore, data centres focus on supporting the tools and data they are offering.[12]
    - Researchers carry out mappings in private scripts and share them in small groups since they are the experts and know what is to be done.
    - Researchers have individual solutions such as spreadsheets or other simple table-like documents[13] which are not meant for public sharing.
- However:
    - Changing semantics is a challenge requiring recording of provenance as well as contextual assertions.
    - In some cases simple mappings between concepts are not sufficient (structural embedding, mapping from simple concepts to composite concepts, etc.).
    - In natural sciences mapping does often include (exact) conversion of units between variables.
    - Any relation between the virtual environments and the real world artefacts should also be recorded.

Life-science disciplines (biomedicine, biodiversity, etc.) are confronted with huge challenges to support mappings between individual concepts and also with aligning complete concept and vocabulary sets to compensate for culture and language dependent differences, in addition to the differences emerging from approaches and theories.

For some time, the life sciences have been well endowed and prepared to invest in semantic infrastructure. The EBI centre, for example, has a semantic group working on and offering semantic tools to support the researchers using their services. It is the only group to our knowledge that provides a tool (OxO) that supports semantic mapping. For selection and visualisation purposes this tool is closely interacting with their ontology lookup service. This is not the place to review this tool in detail but it makes sense to describe some main characteristics that are related with the SEMAF ideas:

- It is meant to support their services, i.e. it was designed as an inhouse tool.
- One of the core design features is to support practical work based on real data sets and not guided by theoretical considerations.

[11] Which we interpret in a broad sense including also cases where the semantics are not (yet) explicit

[12] Note that some data-centers are asked by national science organizations to broaden their scope, for these semantic interoperability is a real issue and are therefore interested in SEMAF

[13] We noted that simple table like documents and spreadsheets are the most common mapping specifications found in many communities

- It was seen to create and manage mappings between concepts and to align vocabularies by a central group.
- It makes use of XML and database technologies which may have some limitations in expressiveness.

We believe that the intentions come very close to what is envisaged by SEMAF, however, a redesign would be required to offer the user-driven, flexible and standards based framework that SEMAF has in mind.

In biodiversity we can see huge efforts by the Global Biodiversity Information Facility (GBIF) and the Biodiversity and Information Standards (TDWG) initiative, for example, to align concept and vocabulary sets, and semantic mapping is of great importance. Yet all tools which are being used are fragmented and there is no commonly usable tool/ framework supporting flexible user driven semantic mapping, though wikibase gains increasing importance as a semantic repository. LifeWatch ERIC operates EcoPortal, a repository of semantic resources for the ecological domain including mappings, and similar resources are potentially offered by other communities.

Similar developments but perhaps at a lesser level of effort can be found in the Archaeology domain that build a Vocabulary Matching Tool[14]. In the Humanities a Data Modeling Environment[15] was developed that also supports semantic mapping creation by experts. CLARIN provides a metadata infrastructure mapping metadata schema to shared concept registries. From Cultural Heritage, there is the X3ML mapping framework [13] offering more powerful mappings between semantics in XML schema and RDF, and is easier to use than XSLT.

Some experts mentioned the use of commercial products as for instance PoolParty[16] for managing semantic relations either within their organisations or in collaboration projects. We note that there is an undecided discussion concerning conditions for the use of commercial, non-free and/or non-open software in relation to the Open Science principles.

SEMAF can learn from this and other work where semantic mapping plays such an important role already.

Vocabulary Matching

Many crosswalks provide for the mapping of metadata elements, but the necessary translation or alignment between the value schemes of the mapped elements should also be taken into account. Because some of the value schemes depend on very large vocabularies where complete alignment is very costly, it would be important to also support partial alignment that would cover >90% of all cases.

[14] https://vmt.ariadne.d4science.org/vmt/vmt-help.html
[15] https://de.dariah.eu/dme
[16] https://www.poolparty.biz

Computational Mappings

Especially in natural sciences, there is a need to map between different variables to carry out normalisations and harmonisations. The methods can reach from unit conversions and simple interpolations to more complex computations[17]. This means that a flexible mapping framework should include a computational algorithm instead of a logical relation. As indicated, in some cases a simple mapping is not seen as being sufficient but a conditional mapping dependent on some context is required. This also requires being able to specify a procedure.

Central vs. Distributed Mappings

As described, many institutions provide semantic mapping (metadata and content data) facilities to augment their services. In all cases known to us these mappings are maintained and curated by the service team who may use automated procedures. This requires measures of validation allowing users to trust the validity of these mappings. These “centrally provided” mappings also require a formalised process to modify the mappings - which in case of large organisations is a slow process. Centrally provided mappings also need to cope with semantic drift over time, which require careful metadata description, provenance tracking, and versioning.

SEMAF, however, is shifting responsibility also to the individual researcher or data manager or a project that wants to carry out a specific analysis on data from across silos. The goal is therefore a pragmatic one driven by the available data and the goals of the analysis. In this case researchers want to test out different types of mappings quickly. SEMAF tools should make it easy to share mappings, which raises the issues of trust and validation. A SEMAF framework needs to offer mechanisms to annotate mappings offered by researchers, for example in terms of usability.. This should not be seen as a formal validation.

SEMAF should focus on supporting the individual usage scenario, but due to its design could also be used by centrally provided mappings, with associated validation marks. Thus, if SEMAF is used by central services, it will have mechanisms to support validation/ annotation, and versioning and provenance tracking. We have seen examples of infrastructures allowing creation and use of mappings first in a local research context, but which can be later published for use by a larger community.

SEMAF Semantic Mapping Framework

Community Interest

From our interviews and personal experience we find there is a common interest in the topic of semantic mapping which was stated by all interview partners. SEMAF-like frameworks will become increasingly important when cross-disciplinary data usage becomes common

[17] For example in the geospatial domain where spatial data is reprojected - sometimes on the fly - by applying an algorithm. Algorithms have been standardised by way of a set of properties.

practice. This trend of cross-disciplinary usage has been mentioned by almost all interview partners. SEMAF-like frameworks would:

- foster the step towards broadly used and standards-based frameworks;
- help to make concept definitions and relations explicit and thus sharable and manageable;
- replace non-standard-based and heterogeneous inhouse solutions;
- have the potential to be used by the individual researchers and make their efforts FAIR;
- help especially the less funded communities to benefit from semantic mapping tools.

The message therefore from these interviews - confirmed by the insight from many other research infrastructures - to EOSC is very clear: it is worth studying the design and development of such a framework. This should first focus on the open design and format specifications and what is needed to integrate existing components, and leave the development of smart tools to interested parties.

Community-Voiced Requirements

The interviews confirmed that key requirements need to be fulfilled for SEMAF-like frameworks to have a chance to be accepted by users. The following points were mentioned in addition to what was already discussed between the SEMAF task-force members:

- simple logical mappings between two concepts are in many cases not sufficient, there need to be mechanisms to:
    - include procedures (conditions, computations, etc.);
    - support for mapping value schemes e.g. (partial) vocabulary alignment;
    - specify hierarchical context of concepts;
    - enable the relation between simple concepts and composite concepts;
- provenance information and contextual assertions need to be included so that users can easily see who has created mappings when and for which purpose and that changes due to changed semantics can be traced;
- there need to be possibilities to add annotations and/ or validation notes to increase trust, which is of great relevance for centrally provided mappings, but which can also be helpful in open frameworks. This may include formal assertions of validity;
- open standards (RDF, OWL, SKOS, etc.) should be supported if appropriate for the requirements to make mappings machine actionable;
- SEMAF should be implementation agnostic and provide support for different mapping processing technologies, including support for transformation recipe descriptions for humans;
- allow for integrating existing mapping infrastructure components where practical e.g. semantic artefacts as vocabularies and ontologies but also mapping specifications and tools;
- visualisation tools need to be included to allow easy navigation in a complex landscape that will emerge and to allow graphically supported operations;
- funding for bootstrapping projects will be essential to get such frameworks started;
- any kind of pertinent legal restraint must be solved.

Perceived Challenges

From the list of requirements it is already evident that coming to an accepted and broadly used framework for semantic mappings is a challenging task.

- SEMAF needs to have the ability to adapt to many different semantic artefacts (from complex ontologies to simple term lists) that are encoded in different formats.
- SEMAF must specify the framework allowing many teams to create the best technology around the specifications which guarantee standard-compliance, FAIRness and thus integration and interoperability.
- SEMAF should be in principle distributed and federative, allowing several centres to maintain service points, however, a registry should be provided that lists the different well-maintained instances. An example of such Registry of Registries (RoR) agreements is the INSPIRE Register Federation[18].
- If the SEMAF framework would become a success, we can expect a proliferation of semantic mappings from many individuals. This implies scalability, proper management, and navigation facilities combined with the possibility of roles for community experts with editorial or curation responsibility.
- Many tools of different sorts would make use of mapping services, requiring a properly defined interface to enable service usage. SEMAF-compliant implementations should be based on a common API specification for harvesting, content exchange, and maintenance.
- SEMAF should allow integration with already existing semantic artefacts and tool components as specified in the requirements.

A specific challenge will be to define a design that would allow for 80% of the use cases and to leave the complex cases to new versions or specialist efforts - as and when required. The specifications w.r.t. permissions need to be such that later extensions can be introduced by others. The rationale behind this has two sides: (1) Using a semantic mapping framework will be new for many researchers and needs thus to be very simple to handle, and (2) tool building should start simple to support quick start-up solutions, allowing the interested community to learn and schemes to be populated.

Conclusions from the Interviews

There is no doubt that cross-disciplinary analysis is still in its infancy except for centres that have teams that can spend the effort to carry out all kinds of transformations and mappings. However, from the interviews and common knowledge it is obvious that there is an expectation that this will change in the future. Thus there is an interest to be prepared for such a change. As a consequence many researchers see the need to not only make data FAIR, but to also include semantic mappings as part of the FAIR universe which can then be shared and reused. Having a unified approach to creating, managing and reusing semantic mappings as intended by SEMAF would add a mechanism that could attract many researchers, although the challenges may not be underestimated.

[18] https://webgate.ec.europa.eu/fpfis/wikis/display/InspireMIG/Registry+federation+requirements

Currently, we can see differences in addressing metadata and content data mappings. This split is caused by the focus of research infrastructures on visibility and the central type of mapping service. With a SEMAF framework in place, there is no need any more for treating them differently which would be a great step forward.

The experience until now, in general, is that semantic interoperability of content data is widely left to the individual researchers who are the experts and thus know what needs to be done to solve their needs. Turning this into a structured and systematic approach that can be used by many implies also a cultural change. This would have to be changed and a framework needs to be well-maintained. The effort to create and manage semantic mappings and make them sharable may not be underestimated. Much outreach and training will be needed to convince people to make the step towards FAIR and shareable solutions.

For EOSC and its follow-ups we see an excellent opportunity to take a leading role to establish clear specifications and a design that could pave the way to such an integrated landscape of service instances fostering FAIR semantics.

SEMAF Requirements

The requirements for SEMAF presented in this chapter result from discussions between members of the SEMAF task force, originating from its expertise and discussing relevant documents. We do especially refer to some papers that have been published recently but do have some overlap: FAIRsFAIR recommendations for the FAIRification of all kinds of semantic artefacts[19], FAIR Semantics recommendation for the FAIRification of semantic artefacts[20] and Recommendations on FAIRification of Vocabularies by S. Cox and colleagues[21].

The analysis of the interviews with the community experts was used to add, corroborate or correct the list of requirements.

We classify and structure the requirements into: Status of Semantic Artefacts, SEMAF Registry and Data Model, (abstract) Infrastructure and Architecture, User Interface Functionality, and Operations and Implementation.

Status of Semantic Artefacts

There is a need for strengthening semantic interoperability solutions in the research data processing landscape.

1. Open Science by Design[22]. We observed that the awareness about Open Science has increased, but that it is mostly interpreted as Open Science by Publishing (OSP). This will not change data science practices, i.e., we expect an increasing awareness about the move towards Open Science by Design (OSD). While OSP delegates all steps of making data science artefacts FAIR only at the end of projects when publications will appear, OSD stresses that the principles of FAIR and Open Science need to be applied as early as possible in the research process. SEMAF expects that relevant semantic artefacts will be FAIR.
2. Mappings as First-class Citizens on the Internet. Sets of mapping relations will be used in various contexts and circumstances. Therefore, it will be necessary that such sets are being treated as first-class citizens on the Internet, i.e., they must be identified as atomic units that encapsulate all relevant information such as context and provenance either as bit sequences or persistent links. This suggests implementing them as FAIR Digital Objects.

[19] FAIR Semantics, Interoperability, and Services ... - FAIRsFAIRwww.fairsfair.eu › node › pdf
[20] https://zenodo.org/record/4314321#.YF 17 KhUM
[21] Guidelines for FAIR Vocabularies. Zenodo. http://doi.org/10.5281/zenodo.4278055
[22] "National Academies of Sciences, Engineering, and Medicine", "Open Science by Design: Realizing a Vision for 21st Century Research", isbn:"978-0-309-47624-9", doi:"10.17226/25116"

Infrastructure, Architecture, & Data model

1. Semantic mappings and crosswalks[23] should be registered in open registries (SEMAF Registry): SEMAF will include as its core a distributed system of federated registries of mappings (relations) that follow the same set of minimal specifications with respect to the relation data-model, grammar, and an API for exchange and accessing information.
2. Crosswalks are sets of mappings. Mappings are relations augmented by appropriate metadata and identified by PIDs. Crosswalks have their own metadata and PID and can be extended by adding new or deleting old mappings into other crosswalks inheriting metadata.
3. All mappings and crosswalks should be versioned (curation, provenance).
4. At first instance researchers can create mappings between “terms” which may represent semantic concepts or values a concept can take. This requires that it must be possible to refer to such terms independently of their contextual embedding. In a variety of cases the contextual embedding may be necessary to enable conditional mappings.
5. The SEMAF registries are required to use a core metadata schema for describing cross-walks and mappings, beyond that core-schema each registry may store additional metadata provided it is (part of) a published schema. This ensures a high level of compatibility between the registries.
6. Crosswalks and mappings and their metadata are published and made harvestable by every SEMAF registry.
7. Crosswalks and mappings once published are ‘persistent’ but related to later versions by provenance and curation.
8. SEMAF registry information should be accessible to both humans (GUI) and machines (API).
9. SEMAF will provide reproducibility of mappings and crosswalks only, beyond that e.g. specific queries the responsibility lies with any SEMAF client application.
10. SEMAF - compliant registries provide registry or repository metadata aligned with community requirements (re3data, FAIR repository metadata requirements) .

User Interface requirements

1. The SEMAF registry UI for creation/ editing/ management of relations needs to be easy to use, which implies that operations need to be as simple[24] as possible without needing to work directly with complex structures such as RDF triples or quadruples.
2. When creating a mapping, it must be possible to visualise the involved target ontologies and their categories to allow easy selections, to create custom groups and relations.
3. It must be possible to browse through and select existing mappings using copy and paste to create new mappings.

[23] Semantic mapping and crosswalks are a special case of mediations, see the Concepts and Definitions Chapter

[24] It is understood that not all aspects of complex mappings may be visualised in a simple way and this should not limit the functionality

4. Smart search/ browse functionality should be offered based on the available metadata and relation specifics e.g. the nature of the relation and its targets.

Machine access requirements

1. Mappings can be automatically tested and validated.
2. All SEMAF information hosted about a registry is harvestable via at least one widely accepted harvesting protocol.
3. All SEMAF registry crosswalks and mappings are openly accessible via an API.
4. Authenticated users can maintain registry entries and mappings via the API

Operational & Content Management Requirements

We need to differentiate between (a) ontologies developed to cover a certain domain of knowledge, (b) semantic mappings that are provided by an authority to serve a community and (c) pragmatic semantic mappings that are provided by individuals or projects to achieve a temporary goal. Domain experts typically focus on more or less comprehensive ontologies to express relevant concepts and their relations, which are based on specific theories and conceptualisations to capture the essentials of the phenomena of their particular research area. It is the task of these individuals or teams to maintain these ontologies and it is known that these efforts seem to be underfinanced. SEMAF does not mean to interfere with these efforts and so should support:

1. Different modes of semantic artefact management and persistence. In various cases institutions or projects maintain “official semantic mappings” between different ontologies to serve their community. These mappings are centrally maintained and need to be backed by certain authority. The SEMAF infrastructure should be used for such mappings, but need to be tagged as being authorised by a specific team. Pragmatic semantic mappings satisfying a specific research goal created by individuals or within projects need to be distinguished from the previous case and thus need to be tagged as being goal-driven mappings. The potential users need to be able to assess the quality and usefulness of these mappings for their goals.
2. The SEMAF registry (and tools) should support different modes of persistence, supporting long-term availability for published shared (sets of) mappings, and shorter (undetermined) life-times for those created for specific purposes in a private domain.
3. Only SEMAF managers/ system administrators should be able to modify/ delete published cross-walks or mappings from a registry in order to keep provenance trails intact.
4. It should be possible to perform bulk imports to bootstrap the registry with existing mappings from specific projects or domains. This would improve acceptance and act as showcases.

5. The bootstrapping phase should come with adapters to frequently used ontologies[25], so that category lists can be extracted, also special integrations for relevant existing mappings can be provided.

Implementation requirements

SEMAF should follow EOSC recommendations for interoperability[26], publication and consider other broad KOS standards specifically:

1. The PID use for crosswalks and mappings must follow the FDO requirements[27,28]where possible.
2. Encoding formats for crosswalks and mappings should, in the first instance, be based on widely adopted web standards such as RDF and JSON-LD, and converge over time to a narrow set of community-adopted implementations. In practice, there will be many valid encodings in the output of the community, ranging from poorly structured or free-text mappings, to structured but non-ideal transformations such as code or XSLT, to well-structured encodings based on for example triples (such as nano-publications[29]), or quadruples.
3. The recommended internal data model for the mappings will be based on quadruples[30] and this may be used as the basis for a universally implementable mapping and crosswalk specification.
4. New implementations should follow a narrower set of specifications and standards for encoding of mappings, limiting future divergence.

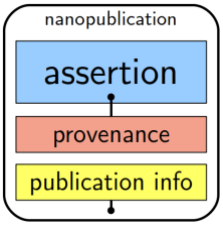
[25] Ontologies can be rather complex and include categories and relations making it hard to extract just the list of categories.
[26] https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf
[27] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF
[28] RDA PID KI. 2019. RDA Recommendation on PID Kernel Information. Research Data Alliance. DOI: https://doi.org/10.15497/RDA00031
[29] http://nanopub.org/guidelines/working_draft/
[30] https://www.w3.org/TR/rdf11-datasets/

Concepts and Definitions

In this document we will use the terms “assertions”, “RDF-triples”,“nano-publications” and “mappings”. The term “RDF-triple” is defined by the RDF definition documents [28]. RDF specifies how a subject and an object can be related by a predicate - all being web resources. As stated in section 3, more recent work tends to use richer formalisms such as property or knowledge graphs, based on n-tuples rather than triples.[31]

“Assertion” is a more generic term taken from logic. It stands for a statement that is true, formulated in some metalanguage.

The notion of “nanopublication” has been introduced recently as the minimum assertion that can be published. It adds typical publication information to an assertion, and is currently formulated as an RDF triple. A nanopublication therefore is technically an assertion formulated in RDF syntax, augmented by well-specified metadata. This currently used binding between nanopublications and RDF syntax could be subject to changes in future.



Assertions such as nanopublications need to exist for long periods of time and need to have stable references over time, i.e. we need to implement them as FAIR Digital Objects.

Next to exact (in a logical sense) terms as “assertion” we also use terms as “(semantic) crosswalk” and “mapping” that are not anchored in logic but rather in practice, and which are used in a less formal way. This still requires us to explain them and their interrelations in more detail. The usage originates from metadata schema crosswalk practice[32], where relations are defined between elements from different metadata schemata in order to facilitate conversions between them. Note that (semantic) crosswalk is also used outside the metadata context (content data), and also in a non-semantic context for managing interoperability on a syntactic or exchange protocol level. See the work done in RDA Brokering Framework WG[33].

The term “mapping” is used for the linking between the concepts used in different descriptive schemata and are created to facilitate interoperability and data conversion. In the context of this document and the proposed SEMAF a semantic crosswalk can consist of many different mappings (see Figure 1 - “SEMAF Registry and Content Model”) and with more detail including the mapping properties in Figure 2 2 - “SEMAF Mapping Model”.

[31] See Glossary for a comprehensive set of terms and definitions
[32] https://en.wikipedia.org/wiki/Schema_crosswalk
[33] https://www.rd-alliance.org/plenaries/rda-17th-plenary-meeting-edinburgh-virtual/brokering-framework-preliminary-recommendations

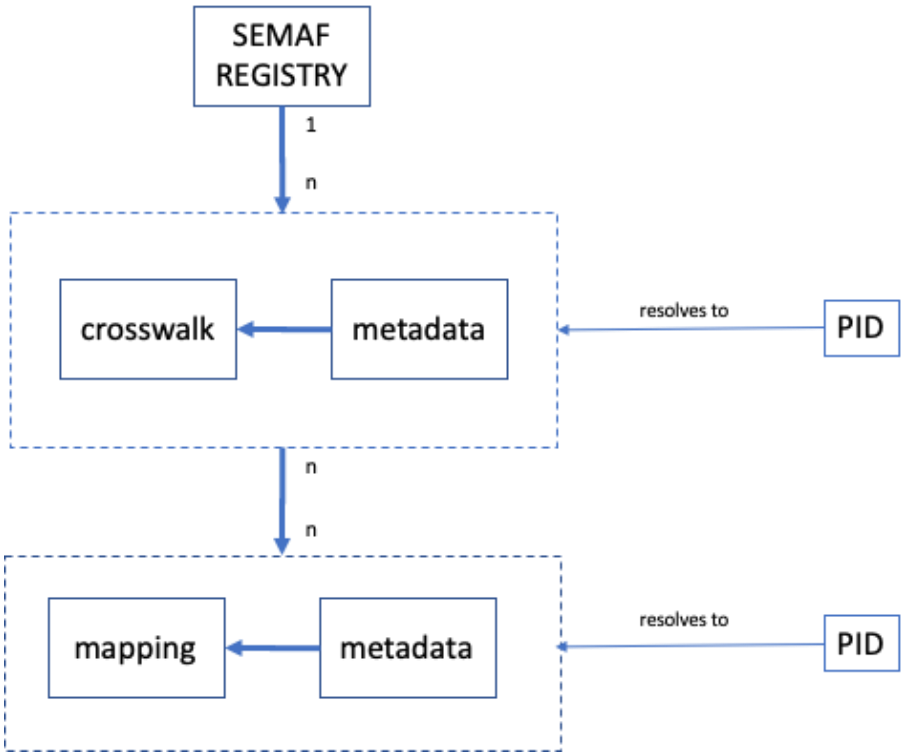


Figure 1. SEMAF Registry and Content Model

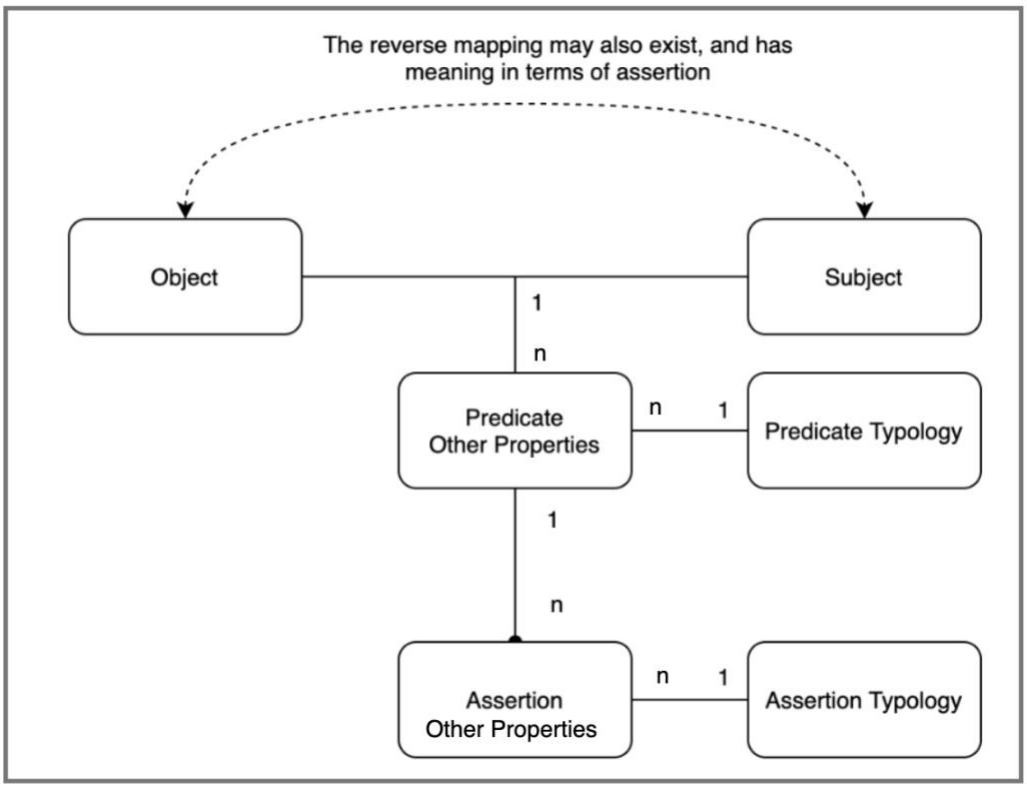


Figure 2: SEMAF Mapping Model

If we look at a more detailed mapping model (Figure 2), the following applies:

1. An object (semantic artefact) will be mapped to another (subject) using a predicate. Since these predicates are all defining some form of mapping (e.g. equivalence or similarity), it will be very useful for machine actionability if these predicates are selected from a community-defined typology.
2. The mapping relation has additional metadata properties (date of creation, creator, etc.).
3. A mapping between the same object and subject (semantic artefacts) can be recorded many times, and the predicates used to define the relation can be, but need not be, the same. Different practitioners may assert divergent relations between the semantic artefacts.
4. There can be more than one assertion associated with the relation defined in the mapping, derived from multiple sources. This applies whether the relation typology is the same for all mappings, and when it is not.
5. Assertions based on a common community-agreed typology will also assist with machine action.
6. Metrics derived from the analysis of multiple mappings (variety of assertion, weight of assertion, assessment of consensus) will be of significant value to the community [38].
7. The same pair of semantic artefacts may be mapped in reverse. In some cases, the predicate is reversible (e.g. 'equivalence'), but in some cases it is not (e.g. isPartOf) but it can still contribute to the metrics about the mapping pair. This implies that information about reversibility will be useful in the predicate typology.

Proposed Architectural Outline

The SEMAF report itself has a limited scope of:

- Providing an inventory of current approaches and investigating the interest and perceived need for a common approach to pragmatic solutions for semantic interoperability via data driven mappings;
- Providing an inventory of current existing components and solutions that can play a role in a future semantic interoperability infrastructure;
- Setting the table for follow-up initiatives and projects, i.e. proposing steps for the implementation of the SEMAF infrastructure, although in the current report it is impossible to cover all aspects, requiring further investigation.

In this chapter we describe several aspects of the overall SEMAF infrastructure, and a number of considerations w.r.t. data model, tooling, and interoperability with existing components and solutions. These particular aspects have come up in discussions of the SEMAF expert task-force itself and/ or in discussions with the community experts that were interviewed. Together they give an overview of how a SEMAF architecture can look without making any decisions on technologies and implementation except, unavoidably, where it concerns interoperability with existing semantic tooling and components.

Federative SEMAF Registry Infrastructure

From a perspective of fostering participation, shared responsibility and also infrastructure robustness, we consider that the infrastructure should support multiple SEMAF registries that collaborate in a federated model. We also consider that in any plans for building SEMAF, there should be initially only one registry that should be fully functional, and able to serve multiple communities and interest groups. The level of integration between the registries in the SEMAF federation can initially be simple - e.g. exchange of mapping metadata only. An extension should be planned to exchange mappings, which would lead to redundant management, improved long-term persistence, and to reliable services. The SEMAF registries will be FAIR from the start, and proper metadata should be provided for the registry as a whole and its published mappings. RDA should be used to define an initial metadata set[34]. Although the level of mandatory collaboration between the different SEMAF registry instances is a matter of the federation contract, we suggest that minimally they should exchange information about the set of SEMAF registry entries and regularly synchronise the mapping metadata. Discovery of the network of SEMAF registries should be achieved by registering it with general semantic KOS registries such as the RDA metadata directory[35], and the different similar discipline specific ones.

Although it is proposed as a self-sufficient independent infrastructure, the SEMAF federation, in turn, conceptually forms part of a wider federation of mediations and transition registries, as foreseen by the Brokering Framework Working Group of the RDA[36]. Work is underway to align the specifications for SEMAF and the Working Group so that such a federation is possible.

Data Model

The data models provided in the requirements section “SEMAF Registry and Content” (Figure 1) and “SEMAF Mapping Model” (Figure 2) seem sufficient for now to capture all requirements but it may need to be refined and extended in a second analysis taking implementation technologies into consideration. We have noted additional requirements for tracking provenance and versioning (“Other Properties” in Figure 2) from our expert interviews. Further choices w.r.t. implementation for the data models and versioning and provenance information should be taken depending on available technical resources and team expertise.

Interoperability and Integration in FAIR Data Landscape

Compliance with requirements for FAIR semantic entities and especially discoverability, SEMAF registry metadata and mapping information should be made available in a number of different encodings and protocols, such as:

[34] See https://www.rd-alliance.org/groups/metadata-ig.html
[35] https://www.rd-alliance.org/groups/metadata-standards-catalog-working-group.html
[36] https://www.rd-alliance.org/plenaries/rda-17th-plenary-meeting-edinburgh-virtual/brokering-framework-preliminary-recommendations

- Registry metadata available as DCMI, DCAT and re3data via OAI-PMH and OpenSearch;
- Suitable SEMAF registry API allowing full access to metadata and content;
- Mapping and crosswalk metadata in a suitable (to be developed) metadata schema, we look to RDA to help find suitable proposals.

Among the SEMAF experts there was consensus to make semantic entities available as FDOs and that the registry API should support selectable return encodings.

Reuse and Integration of Existing Resources

From our interviews and requirements analysis we know there are useful resources already that can be used to seed the registry with little effort. Two options for integrating such existing resources should be supported:

1. Registering only crosswalk metadata and referring to the crosswalk/ mappings available from another site. This enables discovery of such crosswalks via the SEMAF registry.
2. Additionally, also storing the crosswalks themselves after proper conversion to the supported standards.

Note that a bulk import option for existing mappings is considered extremely helpful for seeding the registry, especially for mappings available from existing vocabulary matchings.

Other useful integration resources such as services and tooling should be considered by either developing adapters that allow partial integration with SEMAF infrastructure, or collaboration through their development team for significant integration.

The subject of integrating existing tools and services also make us consider the nature of actual operationalizable mappings. From our inventory they are available in a number of formats:

- Prose text and table based crosswalk recipes for humans[37];
- XSLT type of mapping specifications and similar specifications as in X3ML[38];
- RDF specified mappings;
- Logic built into specific applications.

The intended implementation-agnostic SEMAF infrastructure should manage to handle the first three mapping operationalizations, which we note as accepted SEMAF mapping implementation, but cannot handle mappings implicit in specific tool logic as we noted is often the case for repository submission systems. It is sensible to base any mapping specification on widely adopted standards, such as RDF and JSON-LD, although these may need extensions to accommodate all mapping requirements. Different external technologies to store, organise and access mappings can be supported, facilitating that existing sets of mappings defined by user groups become discoverable and available as FDOs.

[37] E.g. https://www.bgbm.org/tdwg/CODATA/Schema/Mappings/DwCAndExtensions.htm
[38] https://www.ics.forth.gr/isl/x3ml-toolkit

Learning from Early Services & Tooling Examples

The interviews indicated that some communities were already thinking of how to improve data-driven pragmatic semantic mapping which led to a number of fragmented solutions. In particular, we should mention the OxO mapping tool[39] which was developed by EBI and integrated in their services [38] and also the Vocabulary Matching Tool[40] and the Data Modelling Environment[41] from the Humanities discipline that present examples of useful and appealing user interfaces. We need to learn from all these early attempts and need to analyse whether a design and implementation can learn from them and whether an implementation can be built on extending or integrating existing tools.

SEMAF Registry Management Roles and Organizational Embedding

The SEMAF infrastructure design needs to find a suitable compromise between being a reliable source of information and a low-threshold service for developing and extending mappings for specific uses. We find that a clearly visible separation between published and curated content and more incidental mappings including levels of review and approval needs to be supported, but both types of mappings should be supported by SEMAF infrastructure and creation tools. Every SEMAF registry should have a managing organisation, possibly domain specific, curating the published mappings.

SEMAF Proposed Next Steps

This project was focusing on the questions: Do we need such a SEMAF infrastructure? What are the expectations and requirements helping to overcome the current practices - which are far from being FAIR and sustainable? From the survey it was clear that the community would welcome the availability of a flexible semantic mapping framework as proposed by SEMAF.

This chapter makes recommendations about next steps that could help the researchers in changing practices. It needs to formulate answers about specifications of such a SEMAF framework, and how to turn the specifications into a working infrastructure for everyone. It is not meant to repeat all assertions about needs and requirements which have been made in the earlier chapters and recommendations mentioned elsewhere.

This report is based on 26 interviews that have been carried out and on observations about the work in the ESFRI initiatives, in the realm of EOSC discussions, and in RDA.

We will make suggestions of what should be done next mainly based on the requirements from Chapter 4, the interests expressed by the interviewed community experts, and the broad knowledge and experience of the SEMAF task-force members.

[39] https://www.ebi.ac.uk/spot/oxo/
[40] https://vmt.ariadne.d4science.org/vmt/vmt-help.html
[41] https://de.dariah.eu/dme

As next steps we propose initiating the following three phases:

Phase 1: Involvement and Dissemination Phase (I&D)

- The essentials of this report will be discussed with the interviewed experts to see whether their major contributions are represented.
- The essentials of this report need to be made public and be discussed in open forums a) to test the correctness of its conclusions and, if necessary, to make adaptations and extensions, and b) to seek broad support.
- Target of this dissemination and interaction activity are all stakeholders interested but especially the practitioners in the various data and research infrastructure initiatives. This can be done by disseminating the report in the realms of GEDE, RDA and EOSC and by CODATA to reach out to a global audience. Workshops and presentations in meetings will be organised.
- Workshops should also be organised that allow an evaluation of existing tools and practices, discuss a variety of funding models, and determine legal and ethical aspects.
- This phase should not take longer than 6 months to not delay the start of the design and implementation work.
- Some minimal funds should be made available to organise the dissemination and interaction work.

Phase 2: Specification and Design Phase (S&D)

- The final report including its possible adaptations and extensions should lead to formation of a consortium of engaged experts to come to a formal specification and design of a SEMAF infrastructure.
- This group needs to build on what is reported in this document, on what has been recommended elsewhere about FAIRness, TRUST, distributed infrastructure and on tools and mapping resources that already offer interesting options.
- It should be noted that such a SEMAF infrastructure preferably should be organised independent of any single particular academic or economic interests in order to find broad acceptance, i.e., a governance structure and a funding model addressing the sustainability aspects under the EOSC umbrella needs to be defined. With respect to the governance structure it should be noted that a steering board capturing deep research and technology seems to be required. Also this should be in principle be set-up inclusive to allow global participation
- In this phase, a few early adopter initiatives should be actively involved to understand how their semantic artefacts can be integrated and use the semantic mappings providing feedback for design interfaces and bootstrapping. An agile co-design approach is strongly recommended.
- This phase should not take longer than 6 months to not delay the implementation work which could be started partly in parallel following an agile design and implementation style.

Phase 3: Implementation and Test Phase (I&T1, I&T2, I&T3)

I&T Subphase 1

- Based on early specifications and design statements, and on already available practices, tools and resources from some pilot communities, a first prototype should be implemented that supports first simple term-based mappings.
- This first phase implementation needs to
    - also include bootstrapping work allowing interested communities to use the early SEMAF framework to prepare and then upload concept and mapping registrations and
    - allow selected community applications to test and stabilize interfaces.
- A functional prototype should be offered to the public for evaluation after maximally 12 months assuming an agile development process.
- Funding model ideas should be worked out and discussed with the relevant stakeholders.

I&T Subphase 2

- The second phase needs to address the more difficult aspects such as:
    - support for creating complex conditional mappings;
    - fast alignment of vocabularies;
    - smart graphical support;
    - smart search/browse options;
    - deletion options;
    - organisation aspects to support many different user groups
- In this phase also the persistence issue needs to be addressed, i.e., appropriate service providers to host the SEMAF service need to be identified, and the intended funding model needs to be tested.
- Further work in community based bootstrapping and embedding needs to be supported.
- This phase should not take more than 24 months assuming an agile development process.

I&T Subphase 3

- SEMAF needs to become a highly available and reliable service, based on a sustainable funding model, which needs to be in place under the umbrella of EOSC.
- EOSC as the central organisation to foster European data/research infrastructures needs to take responsibility of guiding the future of services such as SEMAF.

Funding & Timing

In the short term, we suggest a fast small funding support action to carry out the dissemination and interaction work immediately after presentation of this report to continue the effort. This should include the explicit goal of the formation of an implementation task-force that can efficiently collaborate with communities that have shown an interest, are already active on the topic, and have resources. We suggest furthermore planning towards funding of a SEMAF followup project for 3 years that takes care of the specification, design, implementation and testing work.

The timing table indicates the months after publication of the SEMAF report.

	1-6	6-12	12-18	18-24	24-30	30-36	36-42
I&D							
S&D							
I&T1							
I&T2							

Terms and Definitions

Term	Definition in SEMAF context
CERIF	Common European Research Information Format
CIDOC-CRM	CIDOC Conceptual Reference Model ontology for cultural heritage conceptual domain, http://www.cidoc-crm.org
CLARIN	European Research Infrastructure for Language Resources and Technology [47]
CMDI	Component Metadata Infrastructure - metadata framework used to describe mainly language data;
concept	Abstract idea conceived in the mind or generalised from particular instances
controlled vocabulary	Closed/ open vocabulary - Set of values that can be used either to constrain the set of permissible values or to provide suggestions for applicable values in a given context [21]
crosswalk	Set of mappings created for solving a specific interoperability challenge and that are managed as a single unit.
data	Refers to any type of bit-sequence that is stored on computers or transferred in networks; in general it is assumed that data contains useful information
DataCite	A widely used schema of data and software metadata, used in conjunction with DOIs [41]
DC (Dublin Core)	A widely adopted metadata schema useful for any digital object [37]
DCAT	A Data Catalogue vocabulary, developed by W3C, that serves as a common definition of elements for many metadata schemata, and can be used as a basis for mapping [40]
DDI	A metadata schema widely used for survey, demography, and social sciences data [42]
Digital Object (DO)	a Digital Object is any digital entity (file, database selection, cloud object, etc.) that has a bitstream (encoded content) and that is associated with a PID and metadata
EOSC	European Open Science Cloud [34]
ERIC	European Research Infrastructure Consortium [36]
FAIR Digital Object (FDO)	a FAIR **DO** is compliant with the FAIR principle and thus machine actionable in all its informational parts including the PID and the metadata. [22], [46]

FAIR, FAIR principles	A set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable [1]
federation of registries	Group of registries that have voluntarily agreed to form a union [26]
framework	Structure of processes and specifications designed to support the accomplishment of a specific task [27]
INSPIRE	Infrastructure for Spatial Information in Europe [43]
interoperability	Ability for two or more systems or applications to exchange information and to mutually use the information that has been exchanged [24]; EOSC interoperability [25]
Link	From hyperlink. Machine actionable reference to a resource.
machine actionability	Any type of data allowing machines to carry out useful processes without human interventions is called “machine actionable” data
mediation	Adapting a digital resource in respect of syntactic, semantic, and/ or schematic interoperability - mediation and adaptation modules are often used to map two different content models or two different interface methods or two different binding types
metadata	In general terms, metadata is data on data; in modern data management we define metadata as the set of properties that are associated with digital objects to enable all kinds of appropriate interpretations and processing
nanopublication	A core scientific statement with associated context. Note 1 In the context of this document a nanopublication is modelled as an “assertion”, “provenance” and “publication info”
ontology	In computer science and information science, an ontology encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many or all domains of discourse. More simply, an ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject [14]
ontology, complex	A complex ontology contains in general a complete description of a domain of knowledge in some formal semantic language, i.e. definition of the categories being used, their properties and their relationships; complex ontologies are associated with particular views on the domain it is describing
ontology, generic	Often a set of categories being used in metadata descriptions is already considered an ontology; a generic ontology is thus anything that includes categories and/ or relationships between categories in some formal semantic language
Open Science, principles of	The OECD defines Open Science as: “to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction” [10]
OWL	OWL Web Ontology Language [15]

persistent identifier, PID	A persistent identifier is a long-lasting ID represented by a string that uniquely identifies a DO and that is intended to be persistently resolved to meaningful state information about the identified DO [29]
provenance information	All metadata categories that are related to the genesis of a specific DO are summarised under the term “provenance”; provenance information can become very complex in data that has been generated by complex workflows. Note also [16], [17]
RDF	RDF — Resource Description Framework [28] - 'RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.'
RDF triples	A semantic triple, or RDF triple or simply triple, is the atomic data entity in the Resource Description Framework (RDF) data model [44]
registry of crosswalks	Registry of crosswalks enables the publication and discovery of a mapping between two semantic artefacts, either at the atomic or collection level
relation	In this context a “relation” specifies the type of relationship between two concepts using well-defined relation types as for example suggested by OWL
research infrastructure	Facilities that provide resources and services for research communities to conduct research and foster innovation
resource	entity, possibly digitally accessible, that can be described in terms of its content and technical properties, referenced by a Uniform Resource Identifier [30]
schema	Formal description of a model [31]
semantic artefact	Machine readable models of knowledge such as controlled vocabularies, thesauri, and ontologies which facilitate the extraction, [linking] and representation of knowledge within data sets using annotations or assertions [6]
semantic crosswalk	A named collection of one or more semantic mappings with a specific purpose
semantic interoperability	Interoperability such that the meaning of the data model within the context of a subject area is understood by the participating systems [18]
semantic mapping	An assertion establishing a relation between two semantic artefacts.
semantic mapping technologies	The specific technology used for encoding or operationalising a specified mapping. e.g. XSLT, Semantic Web technologies, ...
semantic registry	Directory of (authoritative) definitions of terms, concept or data category, or the system maintaining it [19]
semantic space	is a term for the domain of knowledge a set of related semantic artefacts want to cover; the CMDI metadata artefacts are used to describe the nature of possible language resources for example

SKOS	SKOS — Simple Knowledge Organization System [32]
syntactic interoperability	Interoperability such that the formats of the exchanged information can be understood by the participating systems [33]
term	Representation of a semantic concept or value of a semantic concept
Uniform Resource Identifier, URI	A Uniform Resource Identifier (URI) is a unique sequence of characters that identifies a logical or physical resource used by web technologies. URIs may be used to identify anything, including real-world objects, such as people and places, concepts, or information resources such as web pages and books [20]

Bibliography

[1] Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; et al. (15 March 2016). "The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data. 3: 160018. doi:10.1038/sdata.2016.18.

[2] EOSC Interoperability Framework (v1.0) 3 May 2020, https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf

[3] FAIRsFAIR project, https://www.fairsfair.eu

[4] Simon J D Cox, & Alejandra N Gonzalez Beltran. (2020, November). Guidelines for FAIR Vocabularies. Zenodo. http://doi.org/10.5281/zenodo.4278055

[5] Berners-Lee, T.; Hendler, J.; Lassila, O. (2001). "The Semantic Web". Scientific American. 2841

[6] Le Franc, Yann, Parland-von Essen, Jessica, Bonino, Luiz, Lehväslaiho, Heikki, Coen, Gerard, & Staiger, Christine. (2020). D2.2 FAIR Semantics: First recommendations (Version 1.0). FAIRsFAIR, https://doi.org/10.5281/zenodo.3707984

[7] Hugo, Wim, Le Franc, Yann, Coen, Gerard, Parland-von Essen, Jessica, & Bonino, Luiz. (2020). D2.5 FAIR Semantics Recommendations Second Iteration (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.4314320

[8] Floridi, L. (2014). The 4th revolution: How the infosphere is reshaping human reality.

[9] FOSTER Consortium (26 November 2018). "What is Open Science?". Zenodo. doi:10.5281/zenodo.2629946. Retrieved 13 August 2020

[10] OECD (2015). Making Open Science a Reality (OECD Science, Technology and Industry Policy Papers, 25). Paris: OECD Publishing. Available at: http://dx.doi.org/10.1787/5jrs2f963zs1-en

[11] Keith Jeffery, Peter Wittenburg, Larry Lannom, George Strawn, Claudia Biniossek, Dirk Betz, Christophe Blanchi. Not Ready for Convergence in Data Infrastructures. Data Intelligence 3(1), 2021. https://doi.org/10.1162/dint_a_00084.

[12] Main Features of CERIF, https://www.eurocris.org/services/main-features-cerif (retrieved 2021-02-16)

[13] Forth Institute, X3ML, https://www.ics.forth.gr/isl/x3ml-toolkit (retrieved 2021-03-03]

[14] Ontology, Wikipedia, Information Science, (retrieved 10-02-2021) https://en.wikipedia.org/wiki/Information_science

[15] W3C, 2004. Web Ontology Language, http://www.w3.org/TR/owl-guide/

[16] Information technology — Metadata registries (MDR) — Part 7: Metamodel for data set registration, (retrieved 10-03-2021) https://www.iso.org/obp/ui#iso:std:iso-iec:11179:-7:ed-1:v1:en:term:3.1.10

[17] ISO 5127:2017(en), Information and documentation — Foundation and vocabulary, https://www.iso.org/obp/ui#iso:std:iso:5127:ed-2:v1:en:term:3.1.10.26.10

[18] ISO/IEC 21823-1:2019(en)
Internet of things (IoT) — Interoperability for internet of things systems — Part 1: Framework https://www.iso.org/obp/ui#iso:std:iso-iec:21823:-1:ed-1:v1:en:term:3.4

[19] ISO 24622-2:2019(en), Language resource management — Component metadata infrastructure (CMDI) — Part 2: Component metadata specification language. https://www.iso.org/obp/ui#iso:std:iso:24622:-2:ed-1:v1:en:term:3.1.11

[20] Wikipedia, Uniform Resource Identifier, (retrieved 10-03-2021), https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

[21] ISO 24622-2:2019(en) Language resource management — Component metadata infrastructure (CMDI) — Part 2: Component metadata specification language, https://www.iso.org/obp/ui#iso:std:iso:24622:-2:ed-1:v1:en:term:3.1.4

[22] GEDE-RDA-Europe/ GEDE: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF

[23] NISO, Understanding Metadata, (retrieved 10-03-2021), https://www.niso.org/publications/understanding-metadata

[24] ISO/IEC TR 10000-1:1998(en) Information technology — Framework and taxonomy of International Standardized Profiles — Part 1: General principles and documentation framework, https://www.iso.org/obp/ui#iso:std:iso-iec:tr:10000:-1:ed-4:v1:en:term:3.2.1

[25] Oscar Corcho, Magnus Eriksson, Krzysztof Kurowski, Milan Ojsteršek, University of Maribor, Christine Choirat, Mark van de Sanden, Frederik Coppens (2020). EOSC Interoperability Framework (v1.0), Draft for community consultation, https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf

[26] ISO/TR 13128:2012(en) Health Informatics — Clinical document registry federation, https://www.iso.org/obp/ui#iso:std:iso:tr:13128:ed-1:v1:en:term:3.1

[27] ISO/IEC 21823-1:2019(en) Internet of things (IoT) — Interoperability for internet of things systems — Part 1: Framework https://www.iso.org/obp/ui#iso:std:iso-iec:21823:-1:ed-1:v1:en:term:3.3

[28] W3C:, Resource Description Framework (RDF) (retrieved 10-03-2021) https://www.w3.org/RDF/

[29] RDA:, Wittenburg, Helstrom et al. Persistent Identifiers, Consolidated Assertions, Chapter 2 from https://www.rd-alliance.org/system/files/PID-report_v6.1_2017-12-13_final.pdf

[30] ISO 24622-2:2019(en), Language resource management — Component metadata infrastructure (CMDI) — Part 2: Component metadata specification language, https://www.iso.org/obp/ui#iso:std:iso:24622:-2:ed-1:v1:en:term:3.1.10

[31] ISO 19101-1:2014(en), Geographic information — Reference model — Part 1: Fundamentals, https://www.iso.org/obp/ui#iso:std:iso:19101:-1:ed-1:v1:en:term:4.1.34

[32] SKOS Simple Knowledge Organization System Namespace Document - HTML Variant (retrieved 10-03-2021), http://www.w3.org/TR/skos-reference/skos.html

[33] ISO/IEC 19941:2017(en) Information technology — Cloud computing — Interoperability and portability, https://www.iso.org/obp/ui#iso:std:iso-iec:19941:ed-1:v1:en:term:3.1.4

[34] European Open Science Cloud, https://www.eoscsecretariat.eu/

[35] CLARIN, https://www.clarin.eu/

[36] ERIC, https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eric_en

[37] The Dublin Core™ Metadata Initiative, (retrieved 16-03-2021), https://dublincore.org/

[38] Jupp, S., Liener, T., Shantivijai, S., Vrousgou, O., Burdett, T., & Parkinson, H. (2017). OxO - A Gravy of Ontology Mapping Extracts. ICBO.

[39] Naouel Karam, Abderrahmane Khiat, Alsayed Algergawy, Melanie Sattler, Claus Weiland, Marco Schmidt (2020). Matching biodiversity and ecology ontologies: challenges and evaluation results. The Knowledge Engineering Review 35 https://doi.org/10.1017/S0269888920000132

[40] Data Catalogue Vocabulary, W3C. (retrieved 17-03-2021). https://www.w3.org/TR/vocab-dcat-2/

[41] DataCite Metadata Schema 4.3, Released 16 Aug 2019, https://schema.datacite.org/meta/kernel-4.3/

[42] Document, Discover and Interoperate - DDI. https://ddialliance.org/

[43] INSPIRE, https://inspire.ec.europa.eu/

[44] Semantic Triple, Wikipedia, https://en.wikipedia.org/wiki/Semantic_triple

[45] Patricia Harpring, 2010. Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works, Murtha Baca, Series Editor, 2010 J. Paul Getty Trust, ISBN 978-1-60606-026-1 (PDF)

[46] Schwardmann, U., 2020. Digital Objects – FAIR Digital Objects: Which Services Are Required?. *Data Science Journal*, 19(1), p.15. DOI: http://doi.org/10.5334/dsj-2020-015