D5.7 Mapping of technical, repository-specific harvesting endpoints to jointly agreed domain-specific standards and coordination of semantic mapping procedures

Author(s)	Claudia Martens, Anna-Lena Flügel, Jens-Bjørn Riis Andresen
Status	Final
Version	1.0
Date	14. Oktober 2022

Document identifier:	
Deliverable lead	Claudia Martens, Anna-Lena Flügel
Related work package	WP 5
Author(s)	Claudia Martens, Anna-Lena Flügel, Jens-Bjørn Riis Andresen
Contributor(s)	Anders Conrad, Helmut Neukirchen
Due date	November 2022
Actual submission date	
Reviewed by	Debora Testi, Mark van der Sanden
Approved by	
Dissemination level	Public
Website	https://www.eosc-nordic.eu/
Call	H2020-INFRAEOSC-2018-3
Project Number	857652
Start date of Project	1 September 2019
Duration	3 years +
License	Creative Commons CC-BY 4.0
Keywords	Nordic archaeology; semantic mapping; controlled vocabularies; Getty AAT; thesaurus; B2FIND; metadata; findability; discoverability

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Abstract:

This report documents work done in EOSC-Nordic Task 5.1, creating a "Nordic Archaeological" Community in B2FIND and integrating a mapping table that defines different terms for archaeological findings in English, Danish and Norwegian.

Copyright notice: This work is licensed under the Creative Commons CC-BY 4.0 licence. To view a copy of this licence, visit https://creativecommons.org/licenses/by/4.0.

Disclaimer: The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EOSC-Nordic Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EOSC-Nordic Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EOSC-Nordic Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Table of Contents

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Table of Abbreviations

Table 1. Abbreviations appearing in the document

Abbreviation	Explanation
AAT	Art and Architecture Thesaurus
API	Application Programming Interface
B2FIND	EUDAT's interdisciplinary metadata discovery portal
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma-Separated Values
DKRZ	German Climate Computing Centre (Deutsches Klimarechenzentrum)
EOSC	European Open Science Cloud
EUDAT CDI	European Data Collaborative Data Infrastructure
FAIR	Findable, Accessible, Interoperable, Reusable
GIS	Geographic Information System
GUI	Graphical User Interface
JSON	JavaScript Object Notation
LOD	Linked Open Data
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OGC	Open Geospatial Consortium
SLKS	Danish Palaces and Culture Agency (Slots- og Kulturstyrelsen)
SMR	Sites and Monuments Records
WFS	Web Feature Service
WMS	Web Map Service
XML	Extensible Markup Language

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Summary and important takeaways

During the first period of the EOSC-Nordic project, Task 5.1 concentrated on the integration and discoverability of Nordic community-specific data in EOSC via a central search portal: EUDAT B2FIND[1]. This has been described in detail in Deliverable D5.1[2], including lessons learned, a FAIR evaluation and a handbook for metadata ingestion for B2FIND.

In the second period of the project, Task 5.1 focused on a new structure for community-specific 'search spaces' on the one hand and on the implementation of 'semantics' on the other hand in order to enhance both visibility and discoverability of research output even if different languages such as Norwegian, Danish, and English are used. The outcome of these activities is described in this deliverable.

Using an applied rather than a theoretical approach, we investigated how new features in B2FIND could benefit archaeologists in the Nordic countries, namely two national data providers for archaeological research providing sites and monuments records (SMR): the Danish Agency for Culture and Palaces (Slots- og Kulturstyrelsen, or short: SLKS) and the Norwegian Askeladden SMR. These two harvested repositories from Denmark and Norway cover all archaeological finds from these two countries.

Initially, EOSC-Nordic planned to focus only on a subset from these two repositories: finds from the Viking period, as these were considered specifically Nordic[3] – however, this strategy was broadened in the working period, since a restricted focus on the Viking period would have contradicted our long term ambition which aims at general applicability (independent from a particular period). Therefore, we created the "Nordic Archaeology" community (including all records from all periods, i.e. beyond just the Viking period) that may be enlarged by other data providers for archeological research in the Nordic countries. Our starting point are almost 400 000 metadata records (which again are references to millions of observations and findings), making "Nordic Archaeology" the largest B2FIND-community and thus a discovery option serving anybody interested in the remains of the past.

For integrating semantics, we investigated to what extent already existing community-specific thesauri could be re-used or what would be reasons for not using them, respectively. One outcome is a mapping table that allows mapping of different concepts to describe archeological findings used in Norway and Denmark and an additional English translation for these terms. This mapping table is used in B2FIND´s metadata ingestion to enrich values for the `<subject>` field of a record with the assigned terms and hence allows users to search and find archeological data using any language supported by the mapping table.

[1] https://b2find.eudat.eu/
[2] Conrad, A., Martens, C., Flügel, A.-L., Neukirchen, H., Andresen, J., Mihai, H. 2021. D5.1: Discovery and re-use of Nordic community specific data in EOSC, Deliverable, EOSC-Nordic. DOI: 10.5281/zenodo.4607188
[3] See: https://www.eosc-nordic.eu/open-science-will-help-us-better-understand-the-vikings/

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Looking back over the last three project years, some important takeaways from the first deliverable have been confirmed by our subsequent work to be still valid:

- Theoretical concepts for improved interoperability of metadata are extremely important and useful. However, in reality there is a huge gap between theory and practice due to
    - different opinions of what the 'best way' is,
    - legal constraints (on federal/national/European level),
    - ever evolving technologies,
    - not sufficient resources (in terms of human workforce) and
    - political frameworks (for the interaction of already existing national / international / European research infrastructures with new evolving 'european research ecosystems') that more often prevent than enable reliable research infrastructures.
- Standardisation and "FAIRification" of meta/data is an ongoing and mutual process between all partners, which means
    - there is no one-fits-all solution but different ways to ingest and represent metadata,
    - these ways evolve over time and need effort to adapt to changes: a "perfect" solution at the time may be outdated within a year because meanwhile new standards have come up or software was further developed.
- Resources (mainly in terms of staff members) to maintain and update specific solutions (apart from resources for the initial development) should be considered in every project plan. These resources are crucial for sustainable solutions and refer to both
    - human workforce (in order to allow finding appropriate workarounds if a standardised solution is not feasible) and
    - software maintenance (in order to allow the integration of new libraries, new tools, new methods and to guarantee updated configuration of underlying software).

1. Introduction

While large e-infrastructure projects are typically designed to support all aspects of the FAIR principles, actual implementation is not always that simple. A not to be underestimated problem arises from the fact that goals and plans for the technical implementation must already be defined during an e-infrastructure project's proposal process – which does not always fit the current circumstances once the actual implementation starts. Nonetheless, based on the work for the integration of archaeological (meta)data in B2FIND during the first periode of the EOSC-Nordic project, our aim for the second period was twofold: a) to increase search functionalities within the B2FIND search portal by developing a new structure of the underlying CKAN[4] web GUI on the one side, and b) to enhance discoverability of records by implementing community-specific thesauri within the metadata ingestion process on the other side (the reason for using thesauri during the ingestion rather than while the users types in the search is discussed in Section 3.2). Both the conceptual idea behind our developments and the concrete implementation are described below.

[4] https://ckan.org/

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

2. Building the Nordic Archaeology Community

2.1 Starting Point

The starting point for our work has been described in our previous deliverable, i.e. in section "Potential for improvement" of Deliverable D5.1: "One idea is to create a new 'Community' in B2FIND for e.g. 'Nordic archaeology data' and include research data from different repositories (in several countries) within one search interface. It would be useful, if the values in the 'keywords' facet could be filtered according to language settings."[5] This is precisely what we therefore did as a follow-up: a) creating a new community representing metadata records from different data providers and b) integrating a mapping table that aligns different terms in English, Danish, and Norwegian.

The idea behind the "Nordic Archaeology" community was twofold: on the one hand, it should create a common 'search space', enabling the discovery of data within a specific scientific discipline but across national borders. Even though there is a dedicated (European) data infrastructure for archeology and cultural heritage, namely Ariadne plus[6], it is not utilising EOSC technologies and thus does not enable interdisciplinary discovery. Adding to this, the underlying structure should allow the integration of other archaeological data providers in the Nordic countries and elsewhere, which requires a certain flexibility of the ingestion software – which is the case with B2FIND. To implement this flexibility in B2FIND, the ingesting software/pipeline has been revised, rewritten, tested, and deployed.

2.2 Creating the "Nordic Archaeology" Community in B2FIND

As an interdisciplinary search portal, B2FIND a) enables the findability of research data across disciplines, b) enhances the visibility of data providers, and c) acts as a metadata aggregator. All these aspects have been affected by the integration of metadata from SLKS and Askeladden. However, the CKAN structure used internally by EUDAT's B2FIND was only limited to display the value of the 'Communities' field, which is (by the EUDAT Core Metadata Schema[7]) defined to provide the name of "The scientific community, research infrastructure, project or data provider who provides the (meta)data"[8].

[5] Conrad, A., Martens, C., Flügel, A.-L., Neukirchen, H., Andresen, J., Mihai, H. 2021.
[6] https://ariadne-infrastructure.eu/
[7] The EUDAT Core Metadata Schema defines and describes metadata for research output in order to transfer metadata information through different EUDAT CDI services. It originated from the need to define a common schema that allows to harmonise metadata elements used for storage, publication, and discovery of digital research objects across EUDAT partners and beyond. It is build on the Datacite Metadata Schema with additional elements, the XSD file can be accessed here on GitLab: https://gitlab.eudat.eu/eudat-metadata/eudat-core-schema/-/blob/master/eudat-core.xsd, a human readable documentation is available here: https://eudat-b2find.github.io/schema-doc/introduction.html.
[8] https://eudat-b2find.github.io/schema-doc/field_community.html

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

In order to increase search functionalities, we decided to restructure the search portal in a way that allows a broader conceptual idea: B2FIND now displays ‘Communities’ as well as ‘Repositories’ (see Figure 1) where the latter usually refers to the concrete harvesting endpoints while the former may represent discipline specific data providers, infrastructure projects, or other federated systems. Due to the flexible ingestion software, the integration of metadata can work in both ways: one ‘community’ may include metadata from several ‘repositories’ and vice versa, i.e. one ‘repository’ may offer several sets of metadata that belong to different ‘communities’. Both categories are offered as facets and therefore may be used to delimit the search.

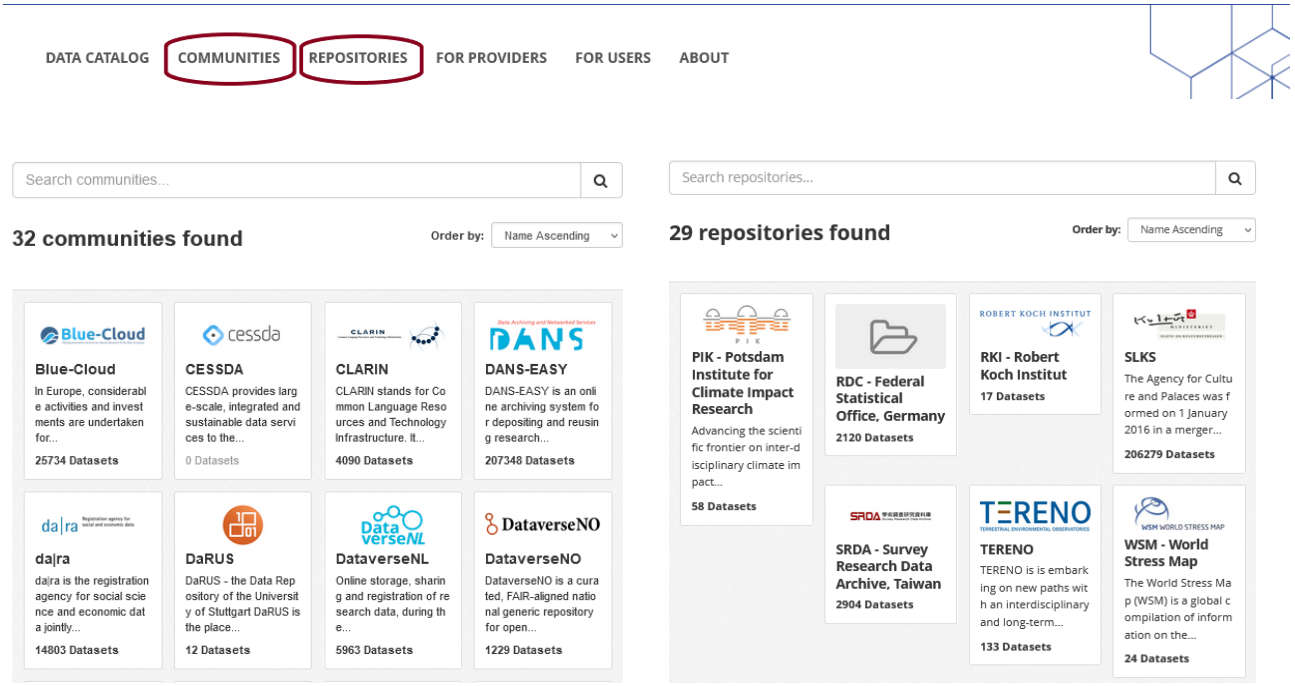

Figure 1. B2FIND user interface at https://b2find.eudat.eu/, now supporting ‘communities’ and ‘repositories (in the top navigation bar, highlighted there using a red frame). By clicking on these, the communities and repositories, shown left-hand and right-hand side, respectively, get displayed. By clicking on them, it is possible to drill down further (see figures 2 and 3). Note that the number of communities and repositories is subject to change, as constantly more and more communities and repositories are added to B2FIND.

For archaeological metadata from Nordic countries, we decided to use the option for a logical concept that encompasses several repositories. Thus we created a new ‘community’ in B2FIND with the label “Nordic Archaeology” that consists of records from both former communities Askeladden (Norway) and SLKS (Denmark). This community has a description for itself as well as a logo and integrated links (which refer to the data providers) as shown in Figure 2.

www.eosc-nordic.eu

EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652

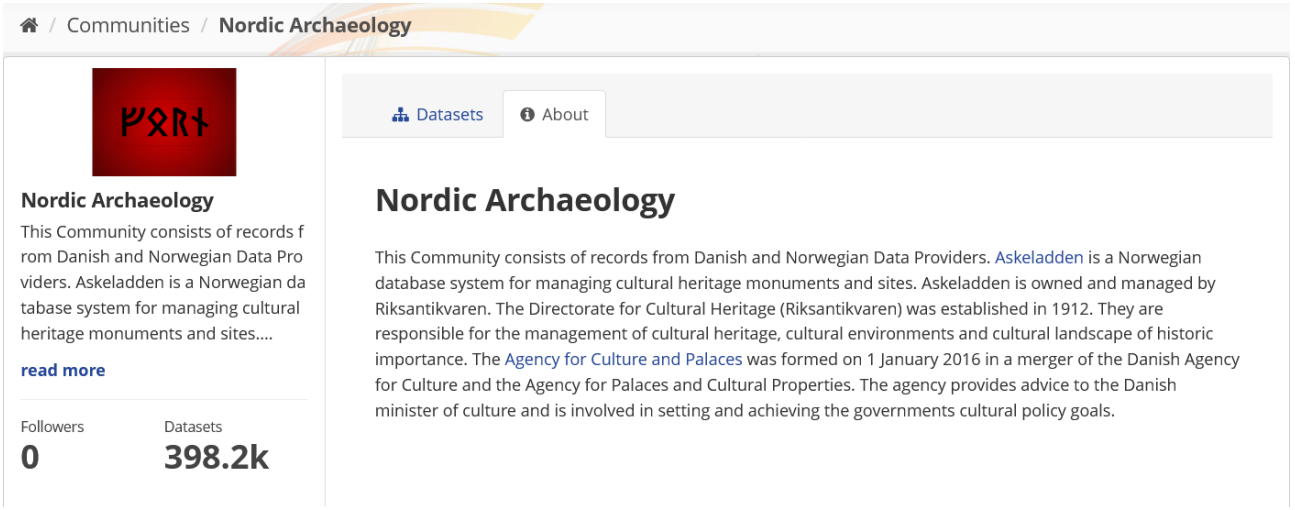

Figure 2. The community "Nordic Archaeology" in B2FIND at https://b2find.eudat.eu/group/nordicar.

While the "Nordic Archaeology" community is represented as an overarching common search space, the individual data providers are visible as well within the section 'Repositories', again with description, logo and integrated links (see Figure 3). In addition, they are listed within the facet 'Repositories' on the search result page.

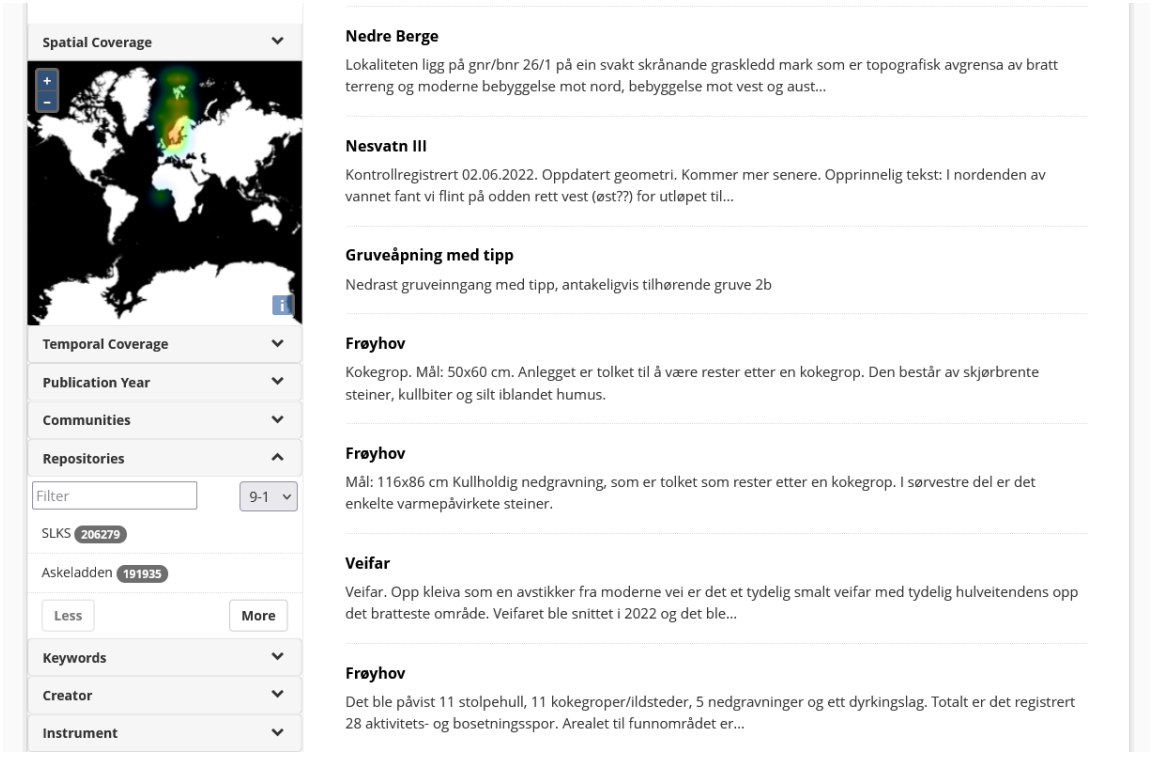

Figure 3. The repository "Askeladden" in B2FIND at https://b2find.eudat.eu/organization/askeladden.

The restructuring of B2FIND´s web GUI required new software development and deployment. It also demanded a complete new ingestion of all existing communities in B2FIND, during which the new logical

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

system had to be applied for all data providers, leading to a revision of B2FIND´s metadata ingestion workflow. Even though this restructuring benefits the B2FIND search portal as a whole (beyond EOSC-Nordic), it was the integration of "Nordic Archaeology" within EOSC-Nordic that triggered this new functionality of B2FIND.

3. Semantics for Nordic Archaeology

"I sympathise with the idea of standards: I think everybody should have one". Prof. Dr. Irwin Scollar – a grand old man within archaeological computing – aired his views in an ironic tone. The time is around 1980-1990: a whole series of nations are moving towards digitising national archaeological records[9]. The *Royal Commision on the Historical Monuments of England* acted as one of the key players, but so too archive repositories in countries from the Continent such as France, the Netherlands, Poland, as well as the Scandinavian countries such as Norway and Denmark, experienced their *digital turn* in those days.

Idiosyncrasies of dated paper records surfaced during the process of digitisation and threatened to reduce the promised potential of digital databases: efficient querying, for instance. How much should one transfer obvious spelling mistakes or outdated language from the paper records to the digital repository? For managers, these problems seemed easy to decide upon and – of course, one would say – thesauri of terms covering all sorts of archaeological finds and contexts were developed and constituted from this moment! The data entered were matched to the standards defined[10].

Every organisation had the power to define their thesauri, which basically were a reflection of the organisational structure and its purpose, professional interests and, maybe, simply coincidences. This meant that the thesauri could be very detailed in some respects – and very coarse in others. In retrospect, a comparative approach reveals differences in structure too. Some thesauri are built up in conceptual hierarchies, from the general to the more specific, whereas others are structured as one consecutive list. Some focus on function, others on material, etc. What is common to all of them is that they rarely change: database managers are not happy with change after the implementation phase.

It was precisely the fear of a loss of dynamics that early critics, such as Torsten Madsen[11], warned us about. His argument was that categories/concepts/terms/vocabularies are conceptual tools developed to cope with complexity, that they are problem-specific and their relevance is subject to change. Given these conditions, the invention and use of standard vocabularies would lead to scientific fossilisation, which on the longer term would result in a less sensible archaeological theory and practice.

However, a more pragmatic approach was taken by Irwin Scollar: his viewpoint was that *"the database was conceived as providing a guide to the sources of information rather than as a replacement for the sources."*[12] Scollar's position was a fair compromise between administrative and scientific goals. At that time the *sources of information* were records – on paper usually – created by the scientific community in academic

[9] Larsen, C.U. (ed.) 1992. Sites and Monuments: National Archaeological Records. The National Museum of Denmark.
[10] See therefore: Davidsen, K. et al. 1979-82. Centralregistrering af stedfæstede fund og fortidsminder. Nationalmuseet.
[11] Madsen, T. 1991. Who said Standardization? CIDOC Newsletter Volume 2, No 1, 20-26;
Madsen, T. 1999. Digital recording of excavations: Do we need data standards and common strategies? Henrik Jarl Hansen & Gillian Quine (eds.) Our Fragile Heritage. Documenting the Past for the Future. København, pp. 131-138.
[12] Scollar, I. 1992. The Bonn Archaeological Database. In: Carsten U. Larsen (ed.) 1992 pp. 92-114

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

freedom, meaning that the author of the sources could assign whatever terms she wanted to the findings she observed.

It is Irwin Scollar's pragmatic approach that we have adopted for B2FIND in the harvesting of the Danish and Norwegian repositories: the meta-data exposed are viewed as a surrogate or a pointer to (further information about) archaeological observations and findings. The ambition of the "Nordic Archaeology" community in B2FIND is meant as a research infrastructure reducing recall and increasing relevance for each specific question one may investigate. The scientific community at least is aware of the fact that information stored in SMR's cannot uncritically be read at face value, and that every single source has to be scrutinised thoroughly before entering the empirical basis of a scientific investigation.

3.1 Work done on Community Side

The sites and monuments records (SMR's) constitute the backbone of archaeology anywhere in the world. Traditionally, one distinguishes between "sites", which are more or less well defined locations fixed to a geographical grid-system and "monuments", which are standing structures accessible for investigation in times of their recording. In some cases, intangible information (myths or legends) associated with a specific location are recorded, too.

The exposed metadata of the harvested repositories summarise how the specific observation or finding is classified using a standard set of terms, in other words: "what it is". These terms come from a closed vocabulary, which is domain-specific and specific for the organisation/body which hosts the repository. It is in this context important to stress that an international classification system, such as the Linnaean system for living organisms[13], does not exist for archaeology. It may well be that the Art and Architecture Thesaurus (AAT)[14] housed by the Getty foundation[15], eventually may develop to a *de facto* agreed vocabulary, since it is a structured vocabulary based on object-oriented principles (class-hierarchies).[16]

In the current version of B2FIND, which does not yet support Linked Open Data (LOD), each participating repository must map a list of harvested metadata (terms) to any other repository. This means that the total number of mappings is N * (N-1) for N participating repositories. Since we raised our level of ambition from a purely Nordic to a European community, we plan that B2FIND in the future will support LOD's. When time comes, a mapping of each (future) repository will therefore only require one mapping, namely to the AAT, which is global in its scope and in its technology. For the two repositories harvested in this use case, this work has already been done, which means that any tag is related to the SPARQL endpoint at http://vocab.getty.edu/. However, none of these currently expose this information via their respective OAI-PMH endpoints that B2FIND harvests from.

Because the terminology of the two repositories are in Danish and Norwegian respectively, each term (the total of terms is around 600) has been translated into English, which again makes it possible for an international audience to query the databases.

[13] Carl Linnaeus published a method to classify living organisms in his "Systema Naturae" in 1735. Since then, the term 'Linnaean system' or 'Linnaean Taxonomy' is used to describe rank-based scientific classifications. See e.g.: Stace, C. A. 1991. Plant Taxonomy and Biosystematics. Cambridge University Press.
[14] Peterson, T. 1990. The Art and Architecture Thesaurus. Oxford University Press.
[15] https://www.getty.edu/research/tools/vocabularies/aat/
[16] A detailed description of the AAT can be found on the website: https://www.getty.edu/research/tools/vocabularies/aat/about.html.

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

3.1.1 SLKS

"Fund og Fortidsminder" (Finds and monuments)[17] contains summary information of about 200.000 archaeological sites and monuments from Denmark and Danish waters. The database has a long history, with roots back to the 17th century[18], and is maintained by the Ministry of Culture. For any site, its history of investigation is listed and for each investigation, there are pointers to archival material. In case of excavation, a full report of the investigation may be accessible for download. A reduced version of the database is accessible for download in a flat file format, whereas the online version is running on top of a relational database, which also contains the classification system in separate tables. This classification system was revised some 40-50 years ago[19] and is a hierarchical system in two layers with top categories such as "Settlement" or "Burial". The total list can be found here: https://docplayer.dk/11776527-Centralregistrering-11-1.html.

3.1.2 Askeladden

"Askeladden"[20] contains summary information of about 200.000 sites with an Individual Protection order from Norway and Norwegian waters[21]. "Askeladden" is focused on Cultural Heritage which is managed by the Ministry of Environment. A reduced version of the database is accessible for the public[22], in addition to web map services (WMS) and web feature services (WFS) to provide map images and map geometry feature vectors (and attributes), respectively, for GIS applications[23]. The categories in the case of Askeladden are subdivided into three classes: one group is their classification in relation to the Cultural Heritage legislation, one is a functional categorization and the third is the type of monument. In summary: Askeladden has a wider chronological scope and a different function than "SLKS", while the latter is more targeted towards archaeology.

3.1.3 Mapping SLKS to Askeladden

The mapping of SLKS to Askeladden that we created is complete, which means that there are no terms in SLKS, which cannot be matched in Askeladden. But it comes at a price, which is lack of precision. Where Askeladden uses one term for burial, SLKS differentiates between 29 types. Where Askeladden uses one term for land tenure boundaries, SLKS differentiates between 24 types. Where Askeladden uses one term

[17] https://www.kulturarv.dk/fundogfortidsminder/
[18] Ebbesen, K. 1985. Fortidsminderegistrering i Danmark. Fredningsstyrelsen. Here: p. 5.
[19] Davidsen et al. 1979-82
[20] https://askeladden.ra.no
[21] Berg, E. 2012. The Use of GIS in the National System for Cultural Heritage Management and Dissemination to the General Public in Norway: Case Study: The Heritage Management Database "Askeladden" and the System for Dissemination to the Public, "Kulturminnesøk". In: Ioannides, M., Fritsch, D., Leissner, J., Davies, R., Remondino, F., Caffo, R. (eds) Progress in Cultural Heritage Preservation. EuroMed 2012. Lecture Notes in Computer Science, vol 7616. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34234-9_59. Here: p.578; see also: https://dokumentasjon.ra.no/askeladden_brukerveiledning/hva_inneholder_askeladden.html
[22] https://www.kulturminnesok.no/
[23] WMS and WFS are specifications of the Open Geospatial Consortium that provide interfaces allowing requests for geographical features across the web using platform-independent calls.

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

for intangible heritage, SLKS differentiates between several types, etc. A search using SLKS terms will thus result in a larger recall and less precision for cases in Askeladden.

3.1.4 Mapping Askeladden to SLKS

The mapping of Askeladden to SLKS that we created is in some respects more problematic. The mapping of terms due to the Cultural Heritage legislation is nearly complete. Because SLKS does not categorise functions and does not include non-archaeological heritage, such as hotels, museums, etc, the mapping of functional terms is almost empty. In addition there are functions that are unknown to the Danish situation, such as reindeer herding and rafting.

A much more satisfactory mapping is the type of monument in Askeladden. Again, if we exclude modern cultural heritage types, such as petrol stations, airports and the like, and if we exclude unknown types as for instance types attached to mountain occupation and use, almost every specific type can find its equivalent.

As in the previous case a search using Askeladden terms will result in a relatively larger recall and less precision in SLKS.

3.2 Work done on Infrastructure Side

3.2.1 Including additional metadata values

Enhancing the interlingual search possibilities and using a research community-specific and widely used controlled vocabulary like the Getty AAT means a great step towards making the metadata FAIRer. The findability is significantly increased by the use of multilingual keywords. The controlled terms of the Getty AAT ensure semantic interoperability with other data and discovery services. But as already mentioned:

- B2FIND does not (yet) support LOD,
- the metadata B2FIND harvests as of now does not contain the Getty global pointers (as the Getty tags are not exposed via the OAI-PMH endpoint at SLKS and the API from Askeladden),
- once it does, B2FIND can **display** the Getty term as additional keywords in English (as English is the preferential language used to search in B2FIND).

However, when it comes to the **search**, different challenges arise:

- The Getty AAT currently lacks Norwegian and Danish translations for most of its terms. B2FIND's goal was to create a searchspace where users can search for archaeological datasets using English, Norwegian and Danish search terms. Implementing the Getty AAT into our Apache Solr[24]-based search scheme would only serve searches with the English Getty AAT term (or other languages available in Getty AAT). Thus it would not be useful for Danish and Norwegian keyword searches.
- B2FIND is an interdisciplinary discovery portal. To serve scientific research communities from all research disciplines alike, we would have to implement several discipline-specific thesauri in our Solr-search. This would result in lots of databases being queried while the users type in their search request. Currently that would slow down the search process considerably. More technical development is needed here.

[24] https://solr.apache.org/

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

To sum up, using the Getty AAT is in theory a good idea, in practice unfortunately not really feasible. Thus, we decided to go another way and implement the mapping table within the metadata ingestion. Even though precision and recall vary with the alignment of specific terms (as described above), it is still better to have *something* instead of *nothing*. Thus while harvesting and mapping metadata from SLKS and Askeladden, additional `<keywords>` are assigned to each JSON record, based on the mapping table created by the archaeological experts. To realise this technically, we added a Jupyter Notebook that enriches the keywords of any given record with their Danish / English / Norwegian equivalents from a translation table, which is a CSV file[25]. As a result, a user may search for, e.g., the keyword “Funerary” and get all records with the Danish keyword “Begravelse” from SLKS and the Norwegian keyword “Gravminne” and vice versa (see Figure 4).

[25] Complementing the FAIR principles, all B2FIND code is openly accessible in Github, including the Jupyter Notebook: https://github.com/EUDAT-B2FIND/md-ingestion

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

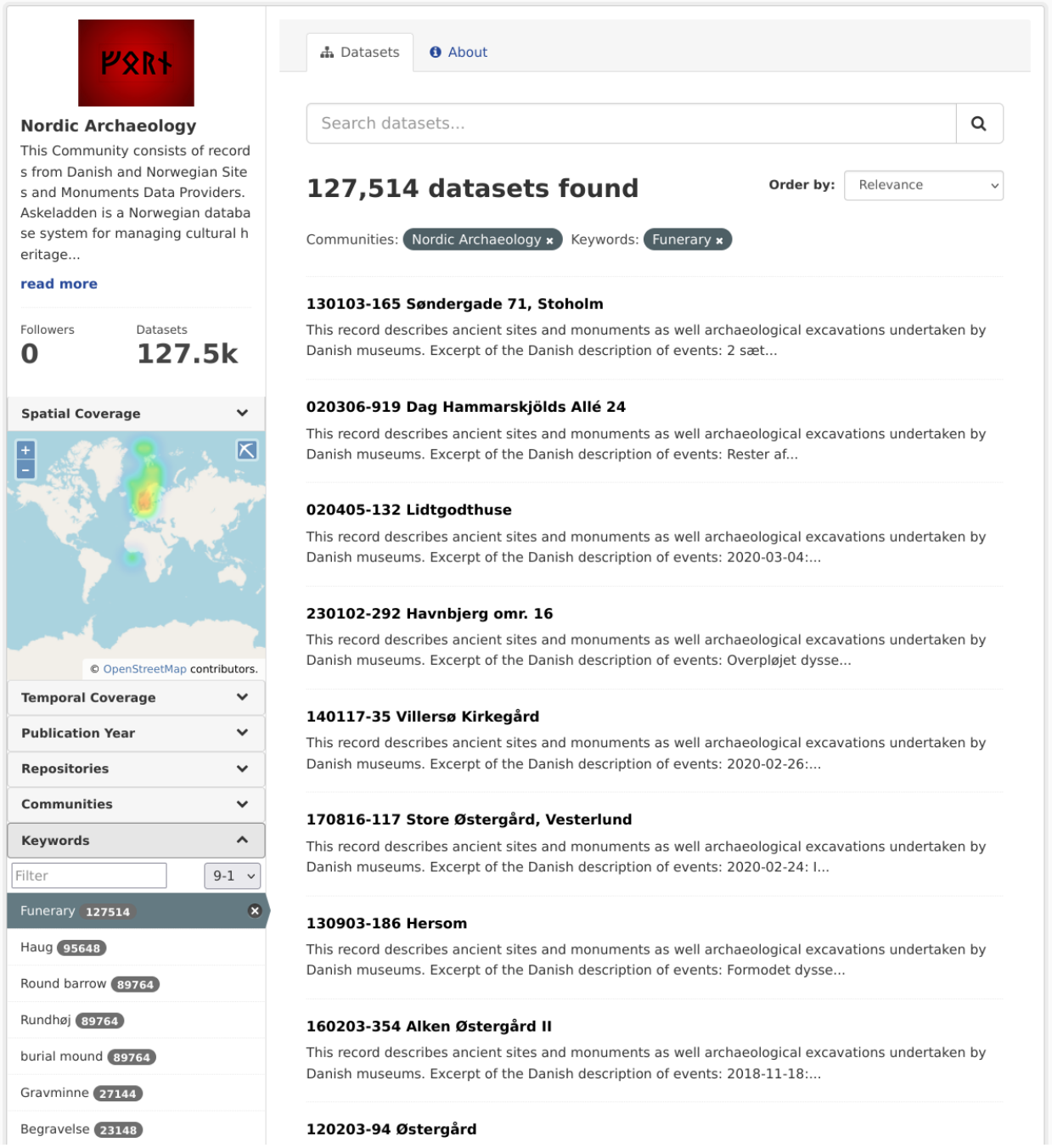

Figure 4. Search in B2FIND for the keyword "Funerary"

3.2.2 Effects for metadata exposure

It should be noted that B2FIND is not only a discovery portal for research output but also a metadata curator, insofar as B2FIND enhances harvested metadata records with additional information. In this case, the originally harvested `<keywords>` have been enhanced with the corresponding translations.

www.eosc-nordic.eu

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

However, B2FIND´s software stack in general allows a flexible mapping as each harvested metadata element may be "updated" within the specific mapfiles[26]. One example is the option to add a default value for `<Discipline>` when records are harvested from a thematic data provider or to add a specific `<contributor>`. Another example is the option to retrieve information for the Boolean operators in `<OpenAccess>` from values in the metadata element `<rights>`.

Apart from being a generic discovery portal and a metadata curator, B2FIND is harvested by OpenAIRE and hence a metadata aggregator as well. Therefore metadata are exposed via a CKAN extension for OAI-PMH (that had to be adapted), using `oai_datacite` as metadataPrefix. Insofar as the internal mapping from EUDAT Core to Datacite allows a match[27], additional information from the ingestion process is exposed. Thus all records from SLKS and Askeladden (including the additional keywords) are also searchable and findable in OpenAIRE Explore[28], broadening the discoverability of Nordic archaeological research even further.

```
−<metadata>
  −<resource xsi:schemaLocation="http://datacite.org/schema/kernel-4 http://schema.datacite.org
   /meta/kernel-4.3/metadata.xsd">
    −<alternateIdentifiers>
      −<alternateIdentifier alternateIdentifierType="URL">
          http://www.kulturarv.dk/fundogfortidsminder/Lokalitet/239041/
        </alternateIdentifier>
      </alternateIdentifiers>
    −<creators>
      −<creator>
          <creatorName>Viborg Museum</creatorName>
        </creator>
      </creators>
      <publisher>Slots- og Kulturstyrelsen (www.slks.dk)</publisher>
      <publicationYear>2020</publicationYear>
      <resourceType resourceTypeGeneral="Other">Dataset</resourceType>
      <language>Danish</language>
    +<titles></titles>
    +<descriptions></descriptions>
    −<subjects>
        <subject>Funerary</subject>
        <subject>Gravminne</subject>
        <subject>Stendyngegrav</subject>
        <subject>Stone heap grave</subject>
        <subject>Archaeology</subject>
      </subjects>
    −<rightsList>
        <rights>Public</rights>
      </rightsList>
    −<geoLocations>
      −<geoLocation>
          <geoLocationPlace>Søndergade 71, Stoholm</geoLocationPlace>
        −<geoLocationPoint>
            <pointLongitude>9.150335</pointLongitude>
            <pointLatitude>56.477965</pointLatitude>
          </geoLocationPoint>
        </geoLocation>
```

Listing 1. XML snippet showing B2FIND´s metadata exposure via OAI-PMH and oai_datacite for a SLKS metadata record, including within the `<subject>` Danish, English, and Norwegian terms.

[26] All mapfiles are openly accessible (currently) in GitHub, here: https://github.com/EUDAT-B2FIND/md-ingestion/tree/master/mdingestion/community

[27] Up to now Datacite does not include a metadata element like `<instrument>`, whereas EUDAT Core does, so this information from B2FIND is not harvested by OpenAIRE. But a mapping for `<temporalCoverage>` is possible, even though a bit tricky.

[28] https://explore.openaire.eu/

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

3.3 Outcome

The outcome of Task 5.1 in EOSC-Nordic is – from our point of view – more than satisfying[29]. It is already implemented and usable in B2FIND. Our results may be summarised as:

- Increased visibility of repositories and their scientific output from the Nordic countries as a result of metadata ingestion in B2FIND and therefore in OpenAIRE, whereas OpenAIRE Explorer will power the EOSC Research Product catalogue.
- Apart from SLKS and Askeladden which have been described in detail here, the following Nordic repositories have been ingested in B2FIND: DATICE[30], DataverseNO[31], and NIRD[32].
- The "Nordic Archaeological" community as a foundation for the integration of more thematic repositories from Nordic countries.
- A mapping table that encompasses different concepts of terms for archaeological findings from Norway and Denmark as a foundation for further terms alignment.
- A common search space for Nordic archaeological data, that can be combined with search results from other disciplines, thus enabling interdisciplinary search for data in one discovery portal.
- The option to use this as a template for other communities being harvested by B2FIND.

While the more intellectual work has been done on the community side, the technical workload on the infrastructure side was enormous. During the first part of the project, the whole B2FIND software stack had to be revised[33] (in order to ensure a certain degree of flexibility), while in the second part, B2FIND´s web GUI has been restructured. One may argue that this development benefits B2FIND and EOSC as a whole, nonetheless it was certainly triggered by the needs of scientific repositories within EOSC-Nordic. Hence regarding software, one of our central lessons learned is that development is important but maintenance is crucial. This applies not only for the software but also for the 'intellectual' outcome of our work, which is the mapping-table for concept-matches from Askeladden and SLKS – if there is a change within the communities regarding these concepts, it is unclear what will happen.

[29] in particular in comparison to the granted PMs.
[30] The Icelandic Social Science Data Service (DATICE) is a data service and archive for Icelandic social science research data, hosted by the The Social Science Research Institute (SSRI). Integrated in B2FIND here: https://b2find.eudat.eu/organization/datice.
[31] DataverseNO is a national generic repository for open research data from all academic disciplines in Norway. Integrated in B2FIND here: https://b2find.eudat.eu/organization/dataverseno.
[32] The NIRD Research Data Archive is a Norwegian repository that provides long-term storage for research data and is compliant with the Open Archival Information System (OAIS) reference model. Integrated in B2FIND here: https://b2find.eudat.eu/organization/nird.
[33] For a precise description, see: Conrad, A., Martens, C., Flügel, A.-L., Neukirchen, H., Andresen, J., Mihai, H. 2021. (cf. footnote 2).

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

4. Outlook

Now as the "Nordic Archaeological" community already exists, it seems desirable to integrate further archaeological data providers in the Nordic countries, such as the Finnish Heritage Agency or the Swedish Rock Art Research Archives. As long as metadata is exposed in a standardised way (using standardised protocols), the integration of new repositories is usual business for B2FIND – but even then, someone has to take responsibility for identifying and connecting possible candidates, and that, however, requires some resources. What happens if there is no standardised way of exposing metadata and a specific repository / community / infrastructure / project needs technical and/or content related support is uncertain. Furthermore, if the "Nordic Archaeological" community should scale up by integrating data repositories from other countries, that would require additional languages mapping and/or concept-matches respectively. So the main question here is how to sustain the work done in the EOSC-Nordic project.

A more concrete outlook refers on the community side to the reuse of archaeological terms translation: submitting the translation work that has been done in this project to Getty for enhancing Getty AAT by Danish and Norwegian terms. On the Infrastructure side, the experience of integrating domain-specific vocabularies into the metadata ingestion process should be extended for other communities. Furthermore, some development is needed to integrate with Solr-based search (e.g. search term recommendations) without slowing the system down. While the communities are ready for Linked Open Data, B2FIND needs to investigate how LOD could be integrated into the search portal (whether for harvesting or for integrating thematic thesauri).

However, even though it is clear which further steps would be useful and desirable, it is up to a higher level of EOSC project management to decide how both sides (scientific community and infrastructure service) could be supported in making these steps.

"Categories are categories of things."

George Lakoff 1987 "Women, Fire, and Dangerous things. What Categories reveal of the Human Mind" p. 9

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652