EOSC NORDIC

DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION

D5.1: Discovery and re-use of Nordic community specific data in EOSC

Author(s)	Anders Conrad, Claudia Martens, Anna-Lena Flügel, Helmut Neukirchen, Jens Andresen, Hannah Mihai
Status	Draft
Version	1.0
Date	25 February 2021

Document identifier:	
Deliverable lead	Anders Conrad
Related work package	WP 5
Author(s)	Anders Conrad, Claudia Martens, Anna-Lena Flügel, Helmut Neukirchen, Jens Andresen, Hannah Mihai
Contributor(s)	Peter Jensen, Ari Lukkarinen
Due date	28 February 2021
Actual submission date	26 February 2021
Reviewed by	Andreas Jaunsen, Troels Rasmussen
Approved by	
Dissemination level	Public
Website	
Call	
Project Number	857652
Start date of Project	1 September 2019
Duration	3 years
License	CC-BY 4.0
Keywords	Nordic, archeology, metadata, harvesting, OAI-PMH, indexing, discovery, FAIR, B2FIND

EOSC NORDIC

DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION

Abstract:

This report documents work done in EOSC-Nordic Task 5.1, harvesting 206.293 datasets from the *Fund og Fortidsminder* collection of records of archeological finds into the EUDAT B2FIND service. Based on how technical challenges were overcome, a more generic "how-to" for harvesting any kind of resource into B2FIND is being presented.

Table of Contents

Table of Abbreviations

Table 1: Abbreviations appearing in the document

Abbreviation	Explanation
API	Application Programming Interface
AU	Aarhus University
CIDOC-CRM	International CommitteeCommittee for Documentation - Conceptual Reference Model
CKAN	Comprehensive Knowledge Archive Network
CMDI	Component MetaData Infrastructure

DDI	Data Documentation Initiative
DDI-CDI	Data Documentation Initiative Cross Domain Integration
DKRZ	German Climate Computing Centre (Deutsches Klimarechenzentrum)
EOSC	European Open Science Cloud
E-RIHS	European Research Infrastructure for Heritage Science
EUDAT CDI	European Data Collaborative Data Infrastructure
FAIR	Findable, Accessible, Interoperable, Reusable
JSON	JavaScript Object Notation
MUSIT	Museum IT, University of Oslo
NeDiMAH	Network for Digital Methods in the Arts and Humanities
netCDF	Network Common Data Form
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ODM	Organization of Danish Museums
OKFN	Open Knowledge Foundation
PARTHENOS	EU supported project: Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
PHP	Hypertext Preprocessor
RDA	Research Data Alliance
RDF	Resource Description Framework
REST	Representational state transfer
SEAD	The Strategic Environmental Archaeology Database
SLKS	Palaces and Culture Agency (Slots- og Kulturstyrelsen)
SMR	Sites and Monuments Records
SPARQL	Query language for RDF
SQL	Structured Query Language
UI	User Interface
URI	Uniform Resource Identifier

UTF-8	Unicode Transformation Format – 8-bit
WFS	Web Feature Service
WMS	Warehouse Management System
XML	Extensible Markup Language
XPATH	XML Path Language
XSD file	XML Schema definition

Introduction

The FAIR principle F4 states that *(Meta)data are registered or indexed in a searchable resource*. Such a searchable resource becomes a kind of exhibit window of the data in question, from which references to the metadata and data should be available. Metadata portals of this kind can be either discipline specific or generic, covering a number of research areas and a multitude of data types, as they do not contain the actual data as such. In the context of EOSC, the EUDAT indexing and search service B2FIND constitutes such a resource, offering indexing of metadata from across disciplines and metadata formats.

In a world where many data resources with their corresponding metadata have been created throughout several years, also before the FAIR principles were published, harvesting and exposing a multitude of data resources from various disciplines can be a challenge. In the context of EOSC it is desirable to streamline this process as much as possible, even as no single standard exists or can be anticipated. The goal of the work reported here, has been to try to create and test a general usable recipe for exposing research data in B2FIND, based on a use case of Nordic archaeological repositories.

For this purpose we have harvested metadata for the Danish *Fund og Fortidsminder* collection into B2FIND. As this is an older resource, it was not a trivial task, but offered technical challenges. How to deal with these challenges in close collaboration between the research community and B2FIND people constituted a major part of the work, resulting in parts of the report being rather technical. Non-technical readers are advised to skim or read through these paragraphs lightly. The level of adaptation to the specific use case performed through this technical work, indicates that the generic “cookbook” that we tried to extract in the last paragraph, can only serve as a starting point.

Other parts of the report cover the B2FIND service and the scientific use case in more general terms. We furthermore report results of using FAIR evaluator tools to explore indications of increased FAIRness of metadata being exposed through B2FIND.

The scientific context of the work is a vision of creating a common Nordic resource for archaeology, which is also being presented. The fulfillment of this vision will be further pursued in future work, which will include adding archaeological metadata from other Nordic collections to B2FIND. Future work could also include harvesting other types of Nordic repositories, aiming to cover geographically across the Nordic countries.

Important takeaways to begin with...

- Theoretical concepts and suggestions for improved interoperability of metadata (in order to enhance discoverability of scientific outcome) are extremely important but cannot always be implemented in practice.
- Standardisation and "FAIRification" of meta/data is an ongoing and mutual process between all partners (standardisation initiatives / projects, scientific communities and data providers) which means
    - there is no one-fits-all solution but different ways to ingest and represent metadata
    - those ways evolve over time and need resources 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 to maintain 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 (that allows to find appropriate workarounds if a standardised solution is not feasible) and
    - software development and maintenance (that allows to integrate new libraries, new tools, new methods and guarantees updated configuration of underlying software).

The B2FIND metadata catalogue and its role in EOSC

Description

B2FIND is a discovery service for research data distributed within EOSC-hub and beyond. It is a basic service of the pan-European data infrastructure EUDAT CDI (Collaborative Data Infrastructure) that currently consists of 29 partners, including the most renowned European data centres and research organisations**.** B2FIND is an essential service of the European Open Science Cloud (EOSC) as it is the central indexing tool for EOSC-hub. Therefore a comprehensive joint metadata catalogue was built up that includes metadata records for data that are stored in various data centres, using different meta/data formats on divergent granularity levels, representing all kinds of scientific output: from huge netCDF files of Climate Modeling outcome to small audio records of Swahili syllables and phonemes; from immigrant panel data in the Netherlands to a paleoenvironment reconstruction from the Mozambique Channel and from an image of “Maison du Chirurgien” in ancient Pompeia to an excel file for concentrations of calcium, magnesium, potassium and natrium in throughfall, litterflow and soil in an Oriental beech forest.

In order to enable this interdisciplinary perspective, different metadata formats, schemas and standards are homogenized on the B2FIND metadata schema, which is based on the DataCite schema extended with the additional elements `<Discipline>` and `<Instrument>`, allowing users to search and find research data across scientific disciplines and research areas as well as searching for certain measurement tools, e.g. data produced by specific beamlines or measurement stations. Good metadata management is guided by FAIR principles, including the establishment of common standards and guidelines for data providers. Hereby a close

cooperation and coordination with scientific communities, Research Infrastructures and other initiatives dealing with metadata standardisation (OpenAire Advance, RDA interest and working groups and the EOSCpilot project to prepare the EOSC including a task on 'Data Interoperability') is essential in order to establish standards that are both reasonable for community-specific needs and usable for enhanced exchangeability. The main question still is how to find a balance between community-specific metadata that serve their needs on the one side and a metadata schema that is sufficiently generic to represent interdisciplinary research data, but at the same time is specific enough to enable a useful search with satisfying search results.

Workflow

B2FIND´s workflow for metadata ingestion basically consists of three steps: harvesting metadata, mapping them and uploading the final JSON records to a database for indexing and search.

Harvesting

Preferably B2FIND uses the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to harvest metadata from data providers. OAI-PMH offers several options that makes it a suitable protocol for harvesting: a) possibility to define diverse metadata prefixes (default is Dublin Core), b) possibility to create subsets for harvesting (useful for large amounts of records or divergent records, e.g. from different projects or sites or measurement stations) and c) the possibility to configure incremental harvesting (which allows to harvest only new/updated records). Nonetheless, other harvesting methods are supported as well, e.g. Open Geospatial Consortium Catalogue Service for the Web (OGC-CSW) or direct JSON-APIs. Harvesting triples from SPARQL endpoints, i.e. Web services supporting the SPARQL query language is implemented only in a beta-version.

Mapping

The mapping process is twofold as it includes a format conversion as well as a semantic mapping based on standardized vocabularies (e.g. the field '`Language`' is mapped on the ISO 639 library and research '`Disciplines`' are mapped on a standardized closed vocabulary). Therefore entries from XML (or JSON) records are being parsed to assign them to the keys specified in the B2FIND schema. Resulting key-value pairs are stored in JSON dictionaries and checked/validated before being uploaded to the B2FIND repository. B2FIND supports generic metadata schemas such as Datacite and Dublin Core. Community-specific metadata schemas are supported as well, e.g. ISO19115/19139 and Inspire for Environmental Research Communities. DDI for Social Sciences in general and CMDI for Linguistics in particular are currently within the frame of a FAIRsFAIR project that aims to improve interdisciplinary research data discovery using DDI-CDI (an "enhanced" version of DDI attempting to make metadata operable across disciplines) and DCATv2 (an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web, also a W3C Recommendation).

Upload and indexing

B2FIND's search portal and GUI is based on the open-source portal software CKAN, which comes with an Apache Lucene SOLR Servlet which allows the indexing of the mapped JSON records and offers performant faceted search functionalities. CKAN was created by the Open Knowledge Foundation (OKFN) and is a widely used data management system. CKAN has a very limited internal metadata schema which has been enhanced for B2FIND by creating additional metadata elements as CKAN field "extra".

B2FIND offers a full text search. Results may be narrowed down using currently 12 facets. These include spatial and temporal search options using a map and an extension for "Timeline Search" and the facets `Communities`, `Keywords`, `Creator`, `Publication Year`, `Discipline`, `Language`, `Publisher`, `Contributor`, `ResourceType` and `OpenAccess`. "`Community`" here is the data provider that B2FIND harvests from. "`OpenAccess`" is "true" per default, if not specified otherwise within `license` or `rights` information.

Integrating archaeological repositories into B2FIND

When describing and evaluating technical solutions for digital research infrastructures, it should be noted that the development, provision and use of software is never a static completed task, but – on the contrary – is subject to continuous development. Particularly in the area of metadata management, technical and content-related requirements are constantly changing; be it through the development and establishment of new standards, be it through new technical interfaces or updated, deprecated or newly available infrastructures. Therefore, the idea of implementing a technical solution that will last forever is a fallacy. The integration of archaeological repositories in B2FIND highlights this major challenge insofar as the whole ingestion code on B2FIND side has been refactored and partly been rewritten during the implementation process, illuminating the fact that software is constantly evolving and in turn affects both the technical implementation and the content structure of the metadata model that is used.

Especially with regard to the importance of FAIR data principles, a fundamental goal is on the one hand to expand the flexibility of (metadata) schemas to ensure suitable findability of data and on the other hand to promote further standardization to increase interoperability.

Starting point – the scientific use case

Culture does not respect modern state boundaries in a globalized world. But also modern history, pre-modern history, and even in prehistory, people travelled, married, traded, followed migrating herds, or had conflicts with their neighbours. Thus new ideas, ways of doing things, materials and moveable artefacts spread from one place to the other – by land, by sea, by air, and recently: by digital technologies.

Since archaeological finds are considered a common resource for humanity, it makes sense to harvest archaeological data from their national repositories and expose them for scientific and other uses. Then, why Nordic? The five Nordic countries – Denmark, Finland, Iceland, Norway, and Sweden (Faeroe Islands, Greenland, and Åland Islands are not considered here) are located in the global top 10 in economic activities related to ICT and are known as European frontrunners in this domain. Stable political systems, high confidence in the state, low corruption rates, and low bureaucratic obstacles in creating new businesses that make the Nordic region an ideal area for new and innovative cross-border activities. A highly educated population and low cultural and low language barriers – at least English as a 2nd language – is an ideal platform for developing new products.

This is why a vision of the unity of all archaeological datasets from all Nordic countries – placed under a European hood, is realistic. These datasets should be accessible by anyone, at any time and serve all thinkable user-groups and open wide avenues for research and other uses.

Despite these similarities of current Nordic societies, their past exhibit enormous variability. This cannot come as a surprise, since the area from arctic North Cape to temperate marshlands in Schleswig, from the harsh North-Atlantic West coast of Iceland to the coniferous east Karelia covers a total over 5.000.000 km^2, the seas including. Thus, the Nordic region shares a cultural past, although not a homogenous one. The first settlers from Norway reached Iceland late in the 9^{th} century AD, whereas the rest of the region was settled from the east as well as from the south after the retreat of Ice Age glaciers. In the course of time, cultures appeared and disappeared spanning sub-regions to varying extents, but always connected through mobility of people and goods – across land or sea.

The cultural variability of the Nordic past is a global representative for any region of this extent and thus again forms an ideal platform for the development of cross-border activities. The untypical of the Nordic region lies more in modern archaeological scholarship. Terms and classification systems – imbedding semantic meaning – are generally shared. This does not imply that there is no disagreement or discussion on these matters, but these discussions always recognize that past cultures did not respect current national borders.

Another rather unusual feature is that most Nordic nations have digital, national data repositories of archaeological findings and that these were established rather early. Digital technologies have been implemented in all major archaeological institutions and in all major workflows. From archaeological fieldwork, over post-excavation analysis and scientific investigations, curating finds and documentation, management and planning, scientific dissemination and activities related to the public in museums or in social media – every step is fundamentally supported by or based on digital technologies.

Also on this background archaeology is commonly acknowledged as the digital frontrunner amongst the humanistic disciplines. Digital methods (typically GIS, databases, statistics) are taught as mandatory classes at universities, specialized courses (typically field recording, finds processing, dissemination) are offered as part of continuing professional development.

Open Science - the vision

The mature digital infrastructures in archaeology combined with the digitally literate workforce are key factors in the growing awareness towards Open Science initiatives in archaeology. In Denmark for instance, the first cultural heritage hackathon (#HACK4DK) was held October 2012 (and has continued annually since), which again inspired the Norwegian #HACK4NO in 2014. There was a strong Nordic presence at NeDiMAH's hackathon November 2011 in London and many other similar events could be listed, e.g. the big Coding da Vinci event.

This is not to say that there still are no mental (i.e. felt "ownership" to archaeological data) and also legal (i.e. copyrights) and political (i.e. non-sharing attitudes) obstacles to overcome, but in general the assertiveness amongst archaeologist towards Open Data is positive. This is reflected in an Open Data initiative by the umbrella organization of Danish museums (ODM). The reason may lie in the fact that much archaeology is publicly funded or financed through legislative procedures. But it may also lie in the tradition in archaeology to have open and collectively shared archives.

For the Information Scientist, archaeology offers a magnitude of possibilities still to be explored (for a recent overview see I. Huvila (2019). Characteristic of archaeological information is its

intrinsic complexity, connectedness and variation of data types. Adding hereto, archaeology produces lots of data. So in terms of quality and quantity archaeology makes a perfect use case for Open Data and Open Science initiatives.

The ultimate goal of uniting all archaeological data of all Nordic countries under FAIR principles and offering the user one search interface, will not be reached overnight. But the perspective to make this treasure accessible for research, education, public planning as well as for other uses, is easily communicated. Enabling searches from a single web interface, the project will increase the accessibility of archaeological and scientific material from the Nordic area. This again will enable collaborations in spite of physical distances, facilitate comparative studies in temporal and spatial terms that have hitherto been limited, and provide research documentation that so far has been available mainly on a local or national level.

The Nordic e-infrastructure will be based on internationally recognized formats and standard protocols like CIDOC-CRM to ensure availability and access. The database system used by the archaeological collections in Norway follows the CIDOC-CRM conceptual model. The database is developed by MUSIT for the Norwegian university museums. Working with systems based on the common CIDOC-CRM ontology will facilitate the integration of the Nordic archaeological data and also make it possible to be linked to resources available through international initiatives like ARIADNEplus, PARTHENOS and E-RIHS.

By connecting large amounts of material with the existing databases, a unique grid infrastructure within the humanities will be created. Inclusion of data through such a Nordic e-infrastructure will substantially impact the archaeological research and the collaboration between archaeology and other disciplines. The planned e-infrastructure will create an arena that initiates new depths of interregional and international research into Nordic prehistory and its palaeo-environment. This approach will also benefit and stimulate collaboration between researchers in the Humanities as in the Sciences. Planned and ongoing research projects will benefit from a Nordic e-infrastructure and the possibility to do online data gathering. Accessing primary material and documentation are often time consuming. It is our opinion that such a Nordic e-infrastructure will facilitate and encourage research approaches leading to new knowledge and new insight.

Nordic Archaeological Sites- and Monuments Records

Fund og Fortidsminder and *Askeladden* contain so-called “sites and monuments records”, shortly SMRs. These repositories refer to findspots of archaeological resources, them being protected ruins of buildings or earthen grave mounds; areas where archaeological investigations have taken place; or find spots of recovered artefacts, socalled “stray finds”. Common to these records are that they possess geospatial information as a property. This is why SMRs in every country are of importance also in physical planning. In order to accommodate the needs by i.e. physical planning in a digital environment, the Danish SMR (*Fund og Fortidsminder*) and the Norwegian SMR (*Askeladden*) offer WMS, WFS, and REST web services.

From a scientific viewpoint, SMRs constitute the backbone of the archaeological information universe, as the SMR links to all other records and documentation, either directly or indirectly. For instance if a scientist studies the isotopic composition of human hair from an archaeological find, that hair is recovered from a body excavated at a specific find-spot in a specific stratigraphic context, so there are links to the excavation report, as well as files on conservation and curation of home taken objects and samples. Thus the Danish SMR was supplied with an OAI-PMH service to cater for the needs of the Europeana initiative.

Of central importance in this context is the role of vocabularies, concepts, ontologies, and standards. Organizations and institutions housing repositories like SMRs often primarily serve management needs and therefore promote standards and standardized concepts in the assignment of archaeological resources. Thus the standards follow the domain of the housing organization, in this case national boundaries. Cross-border initiatives, like EOSC-Nordic, thus have to look into mapping routines and public accessible ontologies to make integration possible.

A further complication is that it is rooted in science to challenge concepts. Without challenging accepted concepts, every scientific field would fossilize. One therefore shall accept that vocabularies/ontologies are dynamic, not static entities.

Technical implementation details

The integration of *Fund og Fortidsminder* records in B2FIND was based on the test ingestion of metadata records from SLKS, the *Danish Agency for Culture and Palaces* (Slots- og Kulturstyrelsen), which took place a few years ago with a specified metadata prefix `<fbb>`. However, the first test ingestion revealed that harvesting from the endpoint 'https://www.kulturarv.dk/repox/OAIHandler' did work, but the metadata prefix had changed to `<ff>`. This non-standard metadata schema had not been defined as a XSD file, did not contain a valid namespace declaration, and could therefore not be validated by the B2FIND ingestion software (because XPATH rules rely on valid namespace declarations). In contrast to `<fbb>`, metadata exposed with the prefix `<ff>` were not used consistently within the namespace. This issue could not be fixed at the source, because changes would have had to happen via the Danish Agency for Culture and Palaces, which reportedly did not maintain the service at the time.

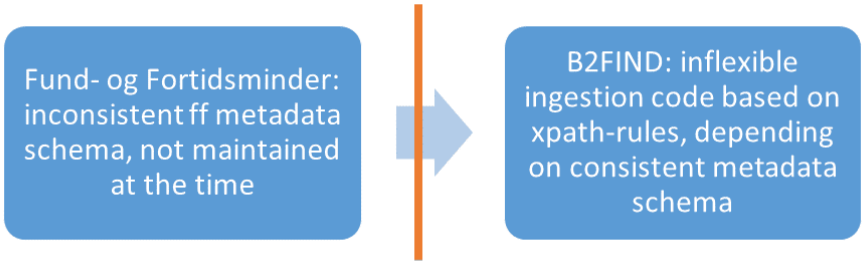

Fig. 1: Technical harvesting fail

In order to show how an ingestion could look like, a demonstrator was developed with some example records and suggestions for the mapping of specific metadata elements.

The solution for integrating all records in B2FIND was to set-up an OAI-PMH server at Aarhus University, where the *Fund og Fortidsminder* partners were in full control of the metadata and could export them in a standardized form, using additionally Dublin Core as metadata prefix. Metadata records from the endpoint 'https://www.archaeo.dk/ff/oai-pmh/' are exposed with both metadata prefixes (`<ff>` as a Community specific metadata schema and `<dc>` as a generic one); for ingesting the records in B2FIND, Dublin Core was used.

Fig. 2: Successful integration of SLKS metadata via staging solution and new B2FIND code

Another issue refers to the use of OAI-PMH “light version” by Aarhus University: B2FIND always displays a link to the originally harvested metadata. This link is built within the ingestion software by using configuration information (endpoint, metadata prefix, oai_subset and oai_identifier), combined with the OAI-PMH command `GetRecord`. As this command is not supported by the “light version” of the Aarhus University OAI-PMH, `MetaDataAccess` could not be automatically generated. In order to ensure a valid `MetaDataAccess`, the developed workaround now links each record in B2FIND to the metadata offered by the SLKS server (and not the one at Aarhus University).

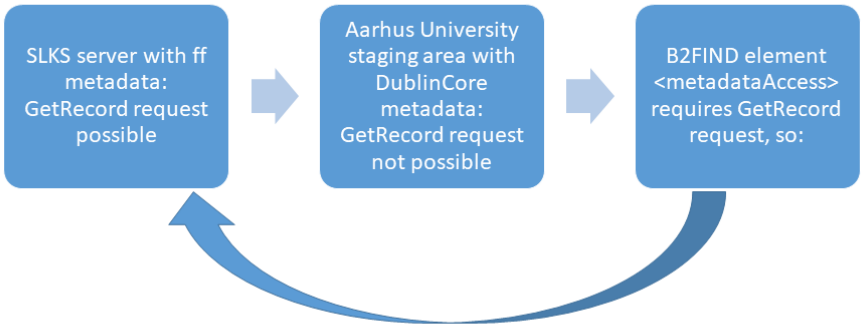

Fig.3: B2FIND <metadataAccess> workaround

Consequently, when viewing a single record, users may click on a link to see all metadata information offered (instead of only those which are mapped to B2FIND schema), even though the information is retrieved from another repository (with `<ff>` metadata schema instead of `<dc>`). This workaround is emphasized in order to highlight that while machine readability of metadata might be the ideal goal of FAIR (meta)data, in reality this aim is not yet achieved. Either the quality of the metadata, its technical provisioning and/or insufficient semantic mapping options between different metadata schemas (especially across disciplines) **often still require a skilled human workforce to define such specific workarounds for each scientific community**. So for the time being, technical interoperability and human insights need to complement each other for good (meta)data management.

Apart from the generic information that is transferred with Dublin Core, the specific mapping for SLKS records contains

a. Discipline = Archaeology
b. PublicationYear = [datestamp in <header> information]
c. TemporalCoverage = [converting “AD” and “BC”]
d. Contact = [Publisher]

Documentation for the integration process of *Fund og Fortidsminder* in B2FIND is important because close communication is key. It is done via a software that allows collective editing and sharing of documents (B2DROP). Additionally, all mapping information is openly accessible on Github.

For B2FIND a “Community” is defined as “The scientific community, research infrastructure, project or data provider from which B2FIND harvests the metadata” and displayed on the webportal with some information. The records from *Fund og Fortidsminder* are integrated within

the Community SLKS, to be seen on the official B2FIND webportal here: 'http://b2find.eudat.eu/group/slks' – another view of the records including the option to use further facets is here: 'http://b2find.eudat.eu/dataset?groups=slks'.

Lessons learned

Exposing metadata

The problems described in the above section "Implementation process", have their root in a faulty configuration of the metadata endpoint that was initially used and could not be changed. Also using the OAI-PMH "light version" required workarounds. Hence, the three most important lessons learned are:

1. always have full control over your metadata…
2. always have full control over your metadata…
3. always have full control over your metadata…

Because of the namespace issues mentioned above, the work was focused on developing a staging area, in which Aarhus University (AU) harvested the relevant data from existing sources (SLKS) through existing APIs (OAI-PMH/REST). This allowed us to verify, augment and correct data before data ingestion into B2FIND.

Subsequently a Python script which iterates through the SLKS *Fund og Fortidsminder* OAI-PMH and reverse engineers the data structure was written. The SLKS delivers records in chunks of 200 items per query, which means iterating through the XML of approximately 900 data-returns.

The XML was re-encoded as SQL statements and ingested into both MySQL and PostgreSQL databases for tests. Supported by a custom UI, developed using PHP, AU could verify that the data structure was consistent and identify data mapping errors in the original dataset and correct these accordingly.

Ingesting metadata

During the integration of metadata records from Fund og Fortidsminder, the B2FIND team identified several deficiencies in the B2FIND ingestion code, which go back to the early stage of B2FIND. Its development started with the onset of EUDAT in 2011, followed up by the project EUDAT 2020 and led to the development of a comprehensive metadata catalogue comprising research data from different scientific disciplines using various metadata schemas and standards. Since the entire complex of metadata management has increased in importance and developed further in recent years, there was a growing need for enhanced modularisation of the core B2FIND ingestion code. Two specific aspects we identified concerning the mapping are:

1. Using XPATH rules to define metadata elements that should be mapped onto a specific target schema is not sufficient because these rules can only be applied if a specific (meta)data model is defined as an XML schema (XSD) with a 'valid' namespace declaration.

2. The B2FIND ingestion software lacked the flexibility that is required for integrating Communities which expose metadata a) in a non standardized way using b) non standardized metadata prefixes.

Based on this insight the decision was made to modify the B2FIND ingestion code. However, what started as 'refactoring' very soon became 'rewriting'. Since the team also successfully insisted on the provision of new hardware, the whole B2FIND system was subject to a fundamental revision. Thus B2FIND now consists of new hardware, a new ingestion code, a new software stack as well as an enhanced metadata schema. Concerning in particular the integration of SLKS and in general the integration of further repositories in the context of EOSC-Nordic, the following features are believed to be of great importance:

- upgrade to new hardware with larger storage space for growing number of records/research communities findable in B2FIND
  = basic pillar for the integration of further repositories, especially regarding those with large amounts of metadata records
- including supplementary libraries in order to reuse existing software (e.g. BeautifulSoup, Shapely)
  = not depending on the usage of XPATH rules
- clean, modular Python-based code in the backend. Enables faster and more flexible metadata integration according to research communities' needs on several complexity levels. Option to group different harvesting endpoints to one Community with specific mapping for each endpoint, including reader for common standards.
  = basic pillar for integrating more 'complex' Communities that have multiple places where meta/data are stored
- enhanced metadata schema based on Datacite, including now the metadata elements `Size`, `Version`, `Funding Reference` and `Instrument`.
  = foundation to display more information and to enable better recall
- exposure of B2FIND metadata records via an OAI-PMH CKAN extension with several metadata prefixes (amongst others: `oai_b2f`, `oai_datacite`)
  = widening the visibility of scientific outcome via further aggregators as e.g. OpenAire

If the B2FIND team had to choose only one 'Lesson Learned' it would be the importance of software development and maintenance. While the whole complex of metadata management refers to FAIR principles (where standardization efforts are fundamental), efforts concerning the underlying technical (development) environment are often neglected.

Assessment of the result

Result - the SLKS community in B2FIND

As of the time of the submission of this report, metadata for more than 200.000 datasets have been harvested from *Fund og Fortidsminder* into the SLKS community in B2FIND. The metadata for the datasets can be searched and displayed via B2FIND's interface. For each dataset, a landing page is available, containing a textual description of the data - in Danish - as well as some structural metadata, including links to the original source of the dataset. Apart from being accessible through the web interface - for humans - the metadata is also available as XML - for machines. The metadata comprise information about provenance and representation, including

spatial coordinates which allows the location of the finding to be shown on a map. Keywords are available both in English and in Danish.
The faceted search interface with the possibility for full text, spatial and temporal search, as well as other search facets, is helpful to narrow down the results. The values in the “keywords” facet are those exposed in the meta-data, and as such are considered important for the specific discipline.

It is not instantly clear to the user how the values within the facets are ordered (default setting is by frequency of occurrence, other orders can be chosen via dropdown menu). At the beginning, only 10 keywords are being displayed. The “More” option for loading the next ten keywords might be overlooked.

The values within the “keywords” facet can be combined using the “+” that appears on the right of the keyword (on mouseover). This is helpful for a researcher who looks for records containing the keywords e.g. “burial mound” AND “Funerary”. This way, search results can be narrowed down using the boolean “AND” conjunction to results containing both keywords. The “OR” operator can be used as well, although only via the freetext search. This is documented in the B2FIND search guide, but not immediately apparent to the user.

A researcher might find it practical to export the results of a search query e.g. in an Excel-file, saving it for later use or for subsequent combination with other searches.

FAIR maturity assessment

The FAIR maturity assessment was performed using the FAIR evaluator tool developed by Wilkinson et al.. This assessment follows a protocol of different checks regarding the (F)indability, (A)ccessibility, (I)nteroperability and (R)eusability of the metadata. The tool was used over the command line and results were collected through an advertised API. More details regarding the FAIR maturity indicator tests can be found directly on the website of the developer of the FAIR

maturity evaluator and in the deliverable of the EOSC-Nordic Work Package 4.

	direct access through	indirect acces through
Source:	kulturarv.dk	B2FIND
Date:	30.11.2020	20.11.20
Dataset	060205-48 Limensgade	060205-48 Limensgade
Collection	6	6
1 F1	success	success
2 F1	failure	failure
3 F1	failure	failure
4 F2	failure	success
5 F2	failure	success
6 F3	failure	failure
7 F3	failure	failure
8 F4	failure	failure
9 A1.1	failure	failure
10 A1.1	success	success
11 A1.2	failure	failure
12 A1.2	success	success
13 A2	failure	failure
14 I1	failure	success
15 I1	failure	success
16 I1	failure	failure
17 I1	failure	failure
18 I2	failure	failure
19 I2	failure	failure
20 I3	failure	success
21 R1.1	failure	failure
22 R1.1	failure	failure

Fig.4: Results of FAIR evaluation. The descriptions of the maturity indicators can be found under the same source as footnote 40.

The evaluation was performed on several random metadata sets from *Fund og Fortidsminder* harvested into B2FIND. To see the advantage of the ingestion of the data into B2FIND, the evaluation was performed on two different access points to the metadata. The first evaluation was performed on the direct access point at 'http://www.kulturarv.dk', the second evaluation was performed over the identifier that B2FIND provides for the dataset. The harvesting into B2FIND improves the FAIR maturity scores (similar to the assignment of DOIs) from 3/22 passed tests to 8/22 (see Fig.4). All tested datasets scored the same, so Fig 4 is representative for all tested datasets. The improvement was made by an automatically generated RDF file during the harvesting into B2FIND.

Testing the same datasets with F-UJI (developed by FAIRsFAIR), a different FAIR evaluator tool the harvesting of the metadata into B2FIND shows no to little improvement of the FAIRness of the metadata. While the evaluator developed by Wilkinson accepts any kind of RDF to pass the tests, the F-UJI tool is a lot more selective. The F-UJI tool is checking the quality of the found metadata in much more detail and it is therefore harder to pass certain tests. This indicates that the automatically generated RDF file is not yet perfect.

It can be expected that the FAIR maturity score of datasets harvested into B2FIND is going to increase in the future (in both, Mark Wilkinsons tool and the F-UJI tool), as the B2FIND team strives to improve the machine readability of the CKAN page sources

It is also worth considering that the F-UJI tool is still under development and the results of the evaluation are based on preliminary code.

Potential for improvements

As already mentioned, the new version of the B2FIND ingestion code allows to integrate different harvesting endpoints (e.g. different subsets within one repository when using OAI-PMH but also different repositories) with specific mappings for each endpoint. Thus, 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. Specifically for this discipline, the time-line search is very important, and calendar dates do not necessarily apply.

Another option is the use of additional keywords (e.g. "nordic archaeological data") for records from SLKS and *Askeladden* in order to enable a search query with sufficient results (as `Keywords` are a facet in B2FIND and thus could be used to narrow down results). In order to get even more satisfactory search results, archaeology-specific ontologies and thesauri could be implemented in B2FIND (on the search side, not on the ingestion side).

Plans for future developments

Including further Nordic repositories into B2FIND

Below is a short list with example candidates of Nordic repositories that could be included in the future. Given the limited resources of EOSC-Nordic, however, it will not be possible to integrate all of them:

- Iceland:
  - GAGNÍS: data from social sciences surveys made on Icelandic population being part of European or international survey programmes EVS (European Values Study), ESS (European Social Survey), or ISSP (International Social Survey Programme), but also from national surveys (such as local elections): 'https://fel.hi.is/is/gagnis-gagnathjonusta-felagsvisinda-islandi'
    - Open datasets are available here: 'https://fel.hi.is/is/gagnis/gogn-i-opnum-adgangi-hja-gagnis'
      While these datasets are currently provided via 'http://idunn.rhi.hi.is/webview/' using the commercial Nesstar WebView software, a migration to using the open source research data repository software Dataverse is currently going on. Given the fact that B2FIND has experience in harvesting from DataverseNO (see below), an ingestion from the upcoming Icelandic Dataverse server is worthwhile.
  - Earth science/Geology data: Given Iceland's geology, a lot of data has been collected in this field, but due to the lack of a data repository in Iceland, this data is typically not available online and thus not yet a candidate for metadata harvesting. As part of other tasks in EOSC-Nordic, the data might be put online.

  - ÍSLEIF archaeological data: While it would fit well into the harvesting of archaeological metadata covered in this document, the data itself is not online and thus not yet a candidate for metadata harvesting. As part of other tasks in EOSC-Nordic, the data might be put online.
- Finland:
  - Etsin -> change in legal issues, planned for first half of 2021
  - Almost all the data provided by Heritage Agency is available on the internet for the public ('https://www.kyppi.fi' – service is so far only in Finnish and Swedish. In that service, one can search archaeological sites on map, get information about archaeological projects and search information from archives and archaeological collections.
- Norway:
  - *Askeladden* (see detailed description above)
  - DataverseNO: already integrated in B2FIND as a Community, here: 'http://b2find.eudat.eu/dataset?groups=dataverseno'
- Sweden:
  - Swedish equivalent to *Askeladden* and *Fund og Fortidsminder*, Fornsök, 'https://www.raa.se/in-english/digital-services/about-fornsok/'
  - Swedish Rock Art Research Archives (https://www.shfa.se)
- Other Nordic interlinked artefact repositories could potentially be included into archaeology community, such as:
  - **Denmark** SARA (based on commercial product) (predecessor is: https://www.kulturarv.dk/mussam)
  - **Norway (in beta test)** Unimusportalen **https://www.unimus.no/portal/#/,**
  - **Iceland** Sarpur (http://sarpur.is)
  - **Sweden** (http://mis.historiska.se/mis/sok/sok.asp?qtype=f), SEAD https://www.sead.se/, Scientific information from archaeological excavations
  - **Finland** Finnish National Library maintains also an open search service entity, FINNA. FINNA contains digital data sets of Finnish libraries, archives and museums, including archaeological material. FINNA supports also international standards, including OAI-PMH. There is an open FINNA API interface available of the finna.fi service

It is envisioned that inclusion of more repository metadata into B2FIND should secure a broader Nordic representation, firstly of archaeological data, but possibly also going beyond that.

Harvesting repository metadata into B2FIND – the cookbook

How-to for harvesting metadata into B2FIND

This section describes the actions a data provider must take in order to publish metadata in the EUDAT-B2FIND catalogue.

1. The first and most important condition for the integration of metadata into the B2FIND portal is that data providers expose their metadata, preferably in a standardised way using

a standardised protocol (such as e.g. OAI-PMH) and a standardised metadata schema (such as e.g. Dublin Core).

2. The data provider must consent to the provided metadata being made publicly available and openly accessible under the Creative Commons Attribution-4.0 International (CC BY-4). The data provider agrees to the metadata being made available for free in B2FIND and also for it to be harvested by and re-distributed to other metadata aggregators under CC BY (e.g. OpenAire). No confidential metadata shall be provided.
3. The next step is to contact the B2FIND team, either via 'http://www.eudat.eu/support-request' or directly (for the duration of EOSC-Nordic) Claudia Martens (martens@dkrz.de) and/or Anna-Lena Flügel (fluegel@dkrz.de).
4. For the providing and mapping of the metadata there are a few mandatory requirements and some good practices or recommendations. For more detailed information, please see the B2FIND guidelines: 'http://b2find.eudat.eu/guidelines/providing.html'. You can find the B2FIND metadata schema with all mandatory, recommended and optional elements here: 'http://b2find.eudat.eu/guidelines/mapping.html'.
5. A test ingestion of the metadata will ensue, accompanied by an exchange of information for preliminary issues, e.g. semantic mapping questions and technical prerequisites. This will be documented in the "Template for Community Integration" (you can find an example here: 'https://b2drop.eudat.eu/s/i9cWHQSb58WoCeC'). Close communication between the B2FIND team and the metadata provider will help to solve all issues and develop community-specific solutions.
6. When both parties are satisfied with the results, the metadata are being ingested into the production B2FIND portal.