Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata

Archana Bidargaddi
Morten Jackobsen
Sikt – Norwegian Agency for Shared Services in E & R

Conference on Smart Metadata for Official Statistics (COSMOS)
11-12 April 2024, Paris
Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata © 2024 by Archana Bidargaddi and Morten Jakobsen is licensed under CC BY-NC-ND 4.0
The Problem with data

F – unable to Find
A – difficult to Access
I – not machine actionable
R – not sufficient metadata
Easier to find and reuse research data

Earlier

Deep metadata is not searchable

Sikt Surveydata

Deep metadata is searchable

Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata © 2024 by Archana Bidargaddi and Morten Jackobsen is licensed under CC BY-NC-ND 4.0
Smart metadata driven

Search
Visualisation
Analysis
Data access

Surveybanken
How concerned are you about climate change?

Example of automatic visualization of answers to the how concerned one is about climate change. The answers are automatically distributed according to background variables defined in metadata such as gender, age and education.

https://surveybanken.sikt.no/en/study/8776be13-4f70-4cc0-8df1-e85803ac3d4d/14?track=analyses&elements=[%7B29e4d471-3314-422f-b0d3-34689a706ba9%7D]&datafile=f26b13a2-553d-4f2b-aeec-22c7cf466c6d/11

Sikt

Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata © 2024 by Archana Bidargaddi and Morten Jackobsen is licensed under CC BY-NC-ND 4.0
How concerned are you about climate change?

Example of automatic map visualisation of answers to the how concerned one is about climate change. Selection of the relevant map to display is driven by metadata.

https://surveybanken.sikt.no/en/study/8776be13-4d70-42c0-8df1-e85803c3ed/2/7?type=analyses&elements=%229eed47f1-3314-422f-b0d3-3e4881a3bba3%22&datafile=f26b538c-553d-47a6-ae62-01bb4846290d/11
How concerned are you about climate change?

Example of own analysis user can perform to explore answers to how concerned one is about climate change. Selection of the relevant visualisation to display is driven by metadata.

https://surveybanken.sikt.no/en/study/8776be13:4d70-4cc0-8df1-e8580-3ed2/1?view=ownAnalyses&elements=[%229eed47f1-3314-422f-bbd1-3489a708a07f%22,%2227e62f-c28e-010d-4eb7-9e1be64c0d85691/2%22]%22&datafile=f26b538c-553d-47af-ace2-011b-4b46290d/11

Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata © 2024 by Archana Bidargaddi and Morten Jackobsen is licensed under CC BY-NC-ND 4.0
Norwegian Research Data Archive

- **Series**: 86
- **Studies**: 2,854
- **Data Files**: 2,853
- **Variables**: 771,351
- **Questions**: 766,973
Background

- 20+ instances of Nesstar
- Over 2500++ studies
- 90% documented at variable level
- Manual data processing in SPSS
- Communication via email
- Redmine ticketing system used for both data processing and data access
- Internal documentation held separately
So then what?
Major infrastructure modernisation projects
From On-premise infrastructure to Cloud Infrastructure

- Old technology
- Disintegrated systems
- Non-collaborative services

To:

- Modern technology
- Integrated systems
- Consolidation of solution
- Cleaning of data and metadata
- Collaborative services
Transforming Static metadata to Smart metadata
Goals

• The curators at the archive should have more control of the metadata they manage
• Shift the authority of the information from the single record to the centralized resource package
• Establish a single source of truth for metadata elements
• Foundation for building metadata-driven systems
Process

- Cleaning and harmonisation of metadata
- DDI-Lifecycle metadata profile
- Transforming metadata to be machine-actionable
- The migration process
- Enabling metadata driven services
- Connecting data and metadata
Metadata quality in legacy system

- DOI in a custom metadata element
- Typos in metadata entry
- Many custom elements
- Manual versioning of the Nesstar-files
- Many instances of free text in place of Controlled Vocabularies and controlled lists
- Subjective tagging av keywords from ELSST
Cleaning and Harmonization

- Information in elements supporting controlled vocabularies were mapped to relevant descriptive terms and code values extracted from CESSDA Vocabulary Service.
- Aggregation to top levels for organisations – from 3000 to 300
- Access conditions mapped to were mapped into 6 broad categories ranging from open access to restricted personal data – from 2700 to 6
- Flatten variable groups
- Dropped all keywords as cleaning would be effort-intensive
DDI-Lifecycle Metadata Profile

Considerations:

• CESSDA Metadata Model had 500 elements
• What was supported by Colectica
• What we had in DDI-Codebook and wanted to carry over
• And what we needed

Results:

• A subset of CESSDA Metadata Model
• Upgrades to Colectica:
  o controlled vocabularies and
  o how the DDI elements Creator, Contributor and Publisher were structured.
Transforming metadata to be machine-actionable

• Defining Key Elements as Resource Packages
  o Concept schemes, Organization schemes, Code lists, Archives, Groups and Other material

• Mapping study elements to the information in the Resource Packages on import.
Metadata Migration – agile and iterative process

- Harmonization mapping files for cleaning
- Automated batch cleaning of Nesstar files
- Mapping elements upon batch import into Colectica
- Review metadata after import
- Finetune "Metadata profile", "Harmonization mapping files" and batch processes based on review
- Repeat export and import
Enabling metadata driven services

Depending on which key metadata element is referenced by the single record, systems have stable information on how to display, make available, and analyze nuanced digital objects.
Enabling metadata driven services

• Versioning of data and documents
  • MAJOR - inclusion of new country data
  • MINOR - Changes in data or metadata that will influence the use of data or the results of data analysis
  • PATCH - insignificant changes such as spelling errors

• Versioning of metadata
  • Every version of an item committed to the repository is saved, allowing clients to retrieve a full version history of any item in the repository.

• Controlled Vocabularies
  • Increased interoperability was achieved by implementation of controlled vocabularies and statistical classifications

• Universal Unique Identifiers
  • Every metadata element has an UUID, enabling systems to identify and retrieve the element.
Connecting data and metadata

- Versioned metadata elements
- Immutable, flat-file for data storage
- Reference data in metadata
- PhysicalInstanceId+version drive data download and analysis
- Data processing APIs to calibrate data with metadata real-time
The new metadata driven Research Data Platform
Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata © 2024 by Archana Bidargaddi and Morten Jackobsen is licensed under CC BY-NC-ND 4.0
Surveybanken

European Social Survey (ESS)

Sikt Data Platform

Data deposit
Data processing and storage
Search engine

Metadata and curation
Statistical engine
Dissemination system

Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata © 2024 by Archana Bidargaddi and Morten Jackobsen is licensed under CC BY-NC-ND 4.0
Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata © 2024 by Archana Bidargaddi and Morten Jackobsen is licensed under CC BY-NC-ND 4.0
Protocols adopted (1)

- **DDI Lifecycle**
  - encourages comprehensive description of data for discovery and analysis
  - facilitates machine-actionability and interoperability

- **GraphQL API**
  - designed to support strong types
  - offers flexibility for various clients consuming GraphQL APIs
  - encourage decoupling of the externally available API and it's internal implementation

- **OIDC/OAuth (for login)**
  - OAuth 2.0 is industry-standard protocol for authorization
  - OIDC is API-friendly
  - OAuth 2.0 capabilities are integrated in the OIDC protocol
Protocols adopted (2)

- Terraform
  - Let's you define infrastructure resources in human-readable configuration files that you can version, reuse, and share
  - Provides a consistent workflow to safely and efficiently provision and manage your infrastructure throughout its lifecycle

- Apache Parquet
  - An open-source columnar data storage file format designed to support fast data access and analysis
  - Compression is performed column by column and supports flexible compression options per data type
  - Optimized for performance and supports data schema evolution

- JSON
  - An open text-based data-interchange format
  - Easy for humans to read and write
  - Machines can easily parse and generate it
Recommendations

- Agile approach, cross functional teams
- API first
- Enabling metadata driven services
- Build data solutions for download and analysis
- Connect data and metadata
Thank you!

archana.bidargaddi@sikt.no
morten.jackobsen@sikt.no